2017

Understanding Public School Accountability Report Design

Michael Moore
University of Pennsylvania, mmoore3@gse.upenn.edu

Understanding Public School Accountability Report Design

Abstract
From the passing of No Child Left Behind (NCLB) through to the signing of the Every Student Succeeds Act (ESSA), all states receiving federal funds must have been required to publicly report on the quality of their schools. In this article, I look at the role of these reports in the broader system of public school accountability, specifically working to untangle the complex interaction between content and form. Drawing on an interdisciplinary body of work in education policy and data visualization and design, I argue that these reports – and more importantly, their visual design – serve as a lynchpin of contemporary school accountability and deserve considerably more attention from both policymakers and practitioners.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Education

First Advisor
Abby Reisman

Subject Categories
Art and Design | Graphic Design

UNDERSTANDING PUBLIC SCHOOL ACCOUNTABILITY REPORT DESIGN

Michael Ryan Moore

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation:

_____

Abby Reisman, Assistant Professor of Education

Graduate Group Chairperson:

_____

J. Matthew Hartley, Professor of Education

Dissertation Committee:

Abby Reisman, Assistant Professor of Education

Janine Remillard, Associate Professor of Education

Rand Quinn, Assistant Professor of Education

UNDERSTANDING PUBLIC SCHOOL ACCOUNTABILITY REPORT DESIGN

DEDICATION

To Nora, who unknowingly gave meaning to this work and motivation to this author.

ACKNOWLEDGMENT

This dissertation exists in large part due to the kindness, thoughtfulness, and support of others. To these people I owe an immense gratitude.

First, I would like to thank the participants in this study. It was a privilege to not only work alongside you, but to hear your stories and your thoughts as we sat and reflected on the work that we had accomplished.

Second, I would like to thank my dissertation committee for guiding me through the often byzantine process of independent research. Thank you to Rand Quinn for pushing me to clarify and strengthen my arguments; to Janine Remillard for appreciating my unique background and allowing me to follow my own path; and finally, to my advisor and chair, Abby Reisman, for enduring my stubbornness, for uncannily organizing my thoughts, and for always advocating for me and my work with passion and earnest excitement.

Third, I would like to thank my many friends and colleagues, particularly David Stewart. None of this would have been possible without David. Six years he took a chance hiring me and has since supported my work unfailingly. Thank you. Furthermore, while nearly every one of my colleagues has supported this work, I would like to especially thank Sam Bishop for his camaraderie and enthusiasm.

In addition, I would like to thank my parents, Karen and Richard Moore. Thank you for reading every single word, not only of this dissertation, but of every academic paper I have written in my career. Words can't express how lucky I am to have you two in my life. I love you both.

Finally, I would like to thank my wife, Meg. Meg, thank you for giving me the courage and strength to tackle this work. Thank you for your patience. Thank you for

your willingness to sit with me and to tackle any problem, no matter how daunting. I am so grateful to have you as my partner.

ABSTRACT

UNDERSTANDING PUBLIC SCHOOL ACCOUNTABILITY REPORT DESIGN

Michael Moore

Abby Reisman

Since No Child Left Behind (NCLB), federal law has required every state to publicly report on the quality of education provided by public schools. These reports, known as school accountability reports, are intended to provide the public with clear and transparent information about school quality so that parents and families may hold those schools accountable, ultimately driving low-performing schools to improve and raising the quality of public education nationwide. Despite serving such a critical role, there is little regulatory oversight on design of these reports. Moreover, there is growing research in the field of data visualization that suggests the design of data reports has deep impacts on the how audiences make sense of, and act on, reported data. This dissertation explores the role of data visualization on these federally mandated school accountability reports.

The first article provides broad context for the issue, detailing the history of school accountability legislation and accountability reports, as well as the relevant research on data visualization, including work specific to education and school reporting. The second article provides a detailed content analyses of several state's current (and historic) accountability reports, looking to understand the status quo of report design, as well as what these design choices suggest about audience and interpretation. Finally, the third article provides an in-depth case study of report design, focusing on how design decisions were made in two different state departments of education, with the goal of

helping practitioners understand how they may shape the design process in their future work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In the spring of 2012, I joined a small consulting firm that specializes in helping public education agencies make sense of education data. At this time nearly all state education agencies (SEAs) and many large urban school districts were regularly collecting a wide array of information about their students, staff, and schools. These education organizations looked to my firm to help them transform these vast stores of raw information into more meaningful insights that could help them to drive decision-making and improve the quality of schools.

As a part of this work, I have had a unique insight into the practice of communicating education data. Though it might seem straightforward, reporting data about students and schools is a politically fraught process, which requires a multitude of nuanced, yet consequential, decisions. These decisions include, but are not limited to: what information to collect; what systems to use in that collection; what measures to calculate; how to perform those calculations; how these measures might incentivize stakeholders (directly or indirectly); and how to present that information to each stakeholder group, recognizing that different audiences engage with information in different ways and that various readers use information in pursuit of various goals.

Recently, my firms' work has focused on solving these problems within the more narrow domain of public school accountability. In 2015, the longstanding accountability legislation No Child Left Behind (NCLB) was replaced by the Every Student Succeeds Act (ESSA). In contrast to NCLB, ESSA provided states with more flexibility in how they defined school quality and, therefore, in how to hold schools accountable for providing a high-quality education. At the same time, ESSA maintained a key NCLB provision requiring states to publish accountability reports, documenting how every single public school in the state measured up to accountability standards. As a result of these

requirements, many states used the passing of ESSA as a chance to rethink their approach to communicating information about schools. Several of these states looked to companies like my own for guidance.

As I started doing this work, I became increasingly aware of the fundamental role that accountability reports play in the broader system of public school accountability. Although states often approached these documents from a compliance mindset, publishing the bare minimum necessary to check a regulatory box, the reports nonetheless served a critical communication function. They were (and remain) the primary mechanism by which the public could assess whether or not any public school, in any state, was meeting its obligation to properly educate students. And yet, despite serving such a critical role, very little attention was paid to these documents. In my experience, state administrators lamented that their reports were not particularly parent- or family-friendly, but viewed report design as a secondary concern to that of report content. Within these departments, teams are hard at work coordinating the massive effort of developing clear accountability models, building stakeholder support for their decisions, collecting and cleaning an enormous amount of student data, and doing all of the above under strict regulatory timelines. While managing this work, states have little capacity to consider the nuances of report design and the (unintended) consequences of micro design decisions. States were more concerned with regulatory compliance than with how the information would make it onto the page. Moreover, even when states did prioritize design, their teams rarely had the capacity or skills to translate good intentions into well designed data visualizations.

Although states' concern with content is understandable, the lack of high-quality design was disheartening. In my work, I have seen time and time again how parents and students are stymied by poorly conceptualized and poorly designed reports. On the other

hand, I have also seen the incredibly positive reaction that parents and students have when they are given reports that take design seriously, reports that understand who their audience is and what their audience is looking for, presenting information for that audience as clearly and accessibly as possible. Furthermore, beyond my day-to-day work with schools, my academic background has also focused extensively on data visualization research. This body of research emphasizes the importance of design choices in facilitating clear understanding and comprehension. Research shows that ill-conceived designs not only frustrate readers, but often lead them to incorrect interpretations of data.

With that in mind, this dissertation represents an early investigation into accountability reports and accountability report design. The dissertation takes a three-article format, with each chapter written as a stand-alone piece, intended for publication separate from the others. As such, some key concepts and research findings are repeated across articles, with the expectation that most readers will not necessarily be reading all three articles in series. The first article provides a broad overview of the topic, exploring the role that accountability reports have played in a changing regulatory landscape, while also introducing the literature on data visualization to argue that closer attention must be paid to the design of accountability reports themselves. Building on this argument, the second article examines the current state of accountability report design. In this article, I conduct a detailed content analysis of multiple states' reports, looking at variation in designs across states, within states, and over time. Finally, the third article looks at how these design decisions are made, providing case studies of two states' attempts to design and disseminate accountability reports. Attention is paid to the design process itself, exploring how various stakeholders advocate for one design over another and, ultimately, how disputes are resolved.

Across all three articles, my goal is to provide practitioners and researchers with a solid ground for understanding the importance of accountability reports as a lever of school accountability and the many ways in which design decisions can affect this lever. Ultimately, I hope that my work serves as a foundation for future research into accountability reports and as a guide for future administrators as they work to provide parents with transparent information about the quality of their schools.

FORM MATTERS: AN ARGUMENT FOR GREATER ATTENTION TO THE DESIGN OF ACCOUNTABILITY REPORTS

Abstract

From the passing of No Child Left Behind (NCLB) through to the signing of the Every Student Succeeds Act (ESSA), all states receiving federal funds must have been required to publicly report on the quality of their schools. In this article, I look at the role of these reports in the broader system of public school accountability, specifically working to untangle the complex interaction between content and form. Drawing on an interdisciplinary body of work in education policy and data visualization and design, I argue that these reports – and more importantly, their visual design – serve as a lynchpin of contemporary school accountability and deserve considerably more attention from both policymakers and practitioners.

Introduction

> There is no such thing as "facts displayed" pure and simple. All facts presented in papers and textbooks are selected from a huge pool of possibilities. Sometimes facts are selected intelligently, appropriate for a purpose; sometimes not – but always they are selected (Macdonald-Ross, 1977, p. 360).

Over the past twenty years, the idea of holding schools accountable for student performance has become a centerpiece of public discourse and political debate (Carnoy, Elmore, & Siskin, 2003; Hanushek & Raymond, 2005; Sirotnik, 2004). And unsurprisingly so. This year, approximately 50 million children will attend public elementary and secondary schools in the United States at an estimated cost of over 600 billion taxpayer dollars (Snyder & Dillow, 2014).

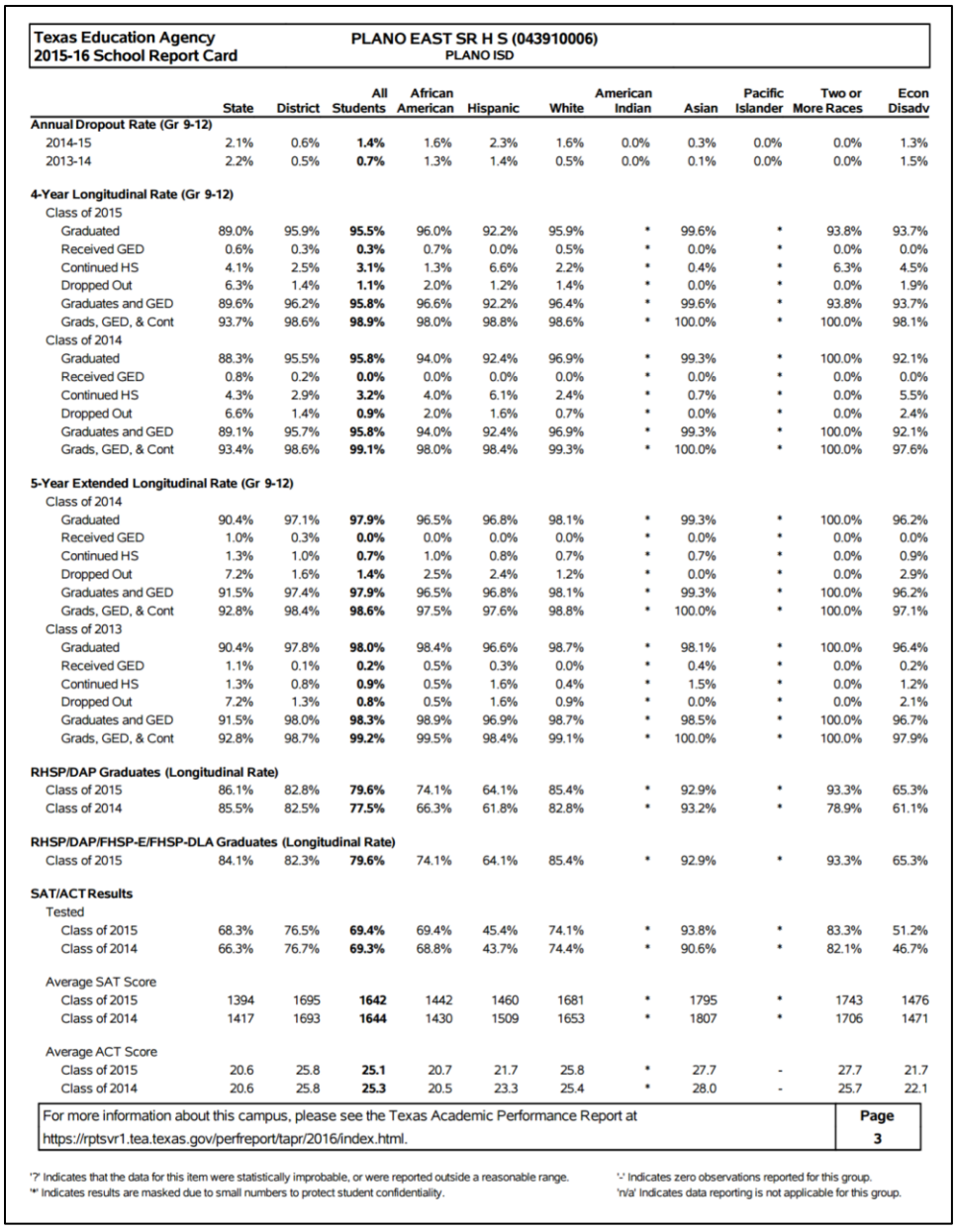
Yet, while many academics have interrogated the politics of holding schools accountable for the public service they provide (Horn, 2002; Hess & Petrilli, 2007; Ravitch, 2010), less attention has been paid to the *communication* of that information. Since the passing of the No Child Left Behind Act of 2001 (NCLB), communicating school accountability has been a key part of accountability legislation. And still today, the Every Student Succeeds Act of 2015 (ESSA) mandates that states publicly disseminate state-, district-, and school-level report cards detailing the quality of their schools via numerous pieces of student performance data (ESSA, 2015).

On the whole, these reports are fairly unassuming documents. Many of the designs seem almost untouched since their first publication. Stern typefaces. Thick gridlines. (See Figures 1 and 2, excerpts from the 2015-16 reports of Texas and New Mexico, respectively). Others take a more visual approach, breaking up the page with bright colors and an array of data visualizations. (See Figure 3, excerpt from the 2015-16 Illinois report). Regardless of format, however, these reports provide readers with detailed performance measures for the schools and their students. On the whole, each report provides a window into the quality of the school as defined by the legislation itself.

This idea that school quality can be represented on a page – and more importantly, that these pages should be made public – is fundamental to the current model of public school accountability. This accountability model suggests that if states provide enough high-quality information about schools to the public, the public will use that information to hold schools to higher standards (Sirotnik, 2004). Power through transparency. But, as it turns out, the way authors put information on the page is quite complex. Although federally mandated reporting may seem one of the dullest of bureaucratic endeavors, it is necessarily political. These accountability reports do not – and cannot – emerge as unbiased documents. Instead, they emerge as the end result of

numerous decisions around what to include and what not to include, what to represent and how to represent it (Winner, 1988). Each of these choices matters.

Figure 1. Sample page from 2015-16 Texas accountability report



Texas Education Agency
2015-16 School Report Card

PLANO EAST SR H S (043910006)
PLANO ISD

	State	District	All Students	African American	Hispanic	White	American Indian	Asian	Pacific Islander	Two or More Races	Econ Disadv
Annual Dropout Rate (Gr 9-12)											
2014-15	2.1%	0.6%	**1.4%**	1.6%	2.3%	1.6%	0.0%	0.3%	0.0%	0.0%	1.3%
2013-14	2.2%	0.5%	**0.7%**	1.3%	1.4%	0.5%	0.0%	0.1%	0.0%	0.0%	1.5%
4-Year Longitudinal Rate (Gr 9-12)											
Class of 2015											
Graduated	89.0%	95.9%	**95.5%**	96.0%	92.2%	95.9%	*	99.6%	*	93.8%	93.7%
Received GED	0.6%	0.3%	**0.3%**	0.7%	0.0%	0.5%	*	0.0%	*	0.0%	0.0%
Continued HS	4.1%	2.5%	**3.1%**	1.3%	6.6%	2.2%	*	0.4%	*	6.3%	4.5%
Dropped Out	6.3%	1.4%	**1.1%**	2.0%	1.2%	1.4%	*	0.0%	*	0.0%	1.9%
Graduates and GED	89.6%	96.2%	**95.8%**	96.6%	92.2%	96.4%	*	99.6%	*	93.8%	93.7%
Grads, GED, & Cont	93.7%	98.6%	**98.9%**	98.0%	98.8%	98.6%	*	100.0%	*	100.0%	98.1%
Class of 2014											
Graduated	88.3%	95.5%	**95.8%**	94.0%	92.4%	96.9%	*	99.3%	*	100.0%	92.1%
Received GED	0.8%	0.2%	**0.0%**	0.0%	0.0%	0.0%	*	0.0%	*	0.0%	0.0%
Continued HS	4.3%	2.9%	**3.2%**	4.0%	6.1%	2.4%	*	0.7%	*	0.0%	5.5%
Dropped Out	6.6%	1.4%	**0.9%**	2.0%	1.6%	0.7%	*	0.0%	*	0.0%	2.4%
Graduates and GED	89.1%	95.7%	**95.8%**	94.0%	92.4%	96.9%	*	99.3%	*	100.0%	92.1%
Grads, GED, & Cont	93.4%	98.6%	**99.1%**	98.0%	98.4%	99.3%	*	100.0%	*	100.0%	97.6%
5-Year Extended Longitudinal Rate (Gr 9-12)											
Class of 2014											
Graduated	90.4%	97.1%	**97.9%**	96.5%	96.8%	98.1%	*	99.3%	*	100.0%	96.2%
Received GED	1.0%	0.3%	**0.0%**	0.0%	0.0%	0.0%	*	0.0%	*	0.0%	0.0%
Continued HS	1.3%	1.0%	**0.7%**	1.0%	0.8%	0.7%	*	0.7%	*	0.0%	0.9%
Dropped Out	7.2%	1.6%	**1.4%**	2.5%	2.4%	1.2%	*	0.0%	*	0.0%	2.9%
Graduates and GED	91.5%	97.4%	**97.9%**	96.5%	96.8%	98.1%	*	99.3%	*	100.0%	96.2%
Grads, GED, & Cont	92.8%	98.4%	**98.6%**	97.5%	97.6%	98.8%	*	100.0%	*	100.0%	97.1%
Class of 2013											
Graduated	90.4%	97.8%	**98.0%**	98.4%	96.6%	98.7%	*	98.1%	*	100.0%	96.4%
Received GED	1.1%	0.1%	**0.2%**	0.5%	0.3%	0.0%	*	0.4%	*	0.0%	0.2%
Continued HS	1.3%	0.8%	**0.9%**	0.5%	1.6%	0.4%	*	1.5%	*	0.0%	1.2%
Dropped Out	7.2%	1.3%	**0.8%**	0.5%	1.6%	0.9%	*	0.0%	*	0.0%	2.1%
Graduates and GED	91.5%	98.0%	**98.3%**	98.9%	96.9%	98.7%	*	98.5%	*	100.0%	96.7%
Grads, GED, & Cont	92.8%	98.7%	**99.2%**	99.5%	98.4%	99.1%	*	100.0%	*	100.0%	97.9%
RHSP/DAP Graduates (Longitudinal Rate)											
Class of 2015	86.1%	82.8%	**79.6%**	74.1%	64.1%	85.4%	*	92.9%	*	93.3%	65.3%
Class of 2014	85.5%	82.5%	**77.5%**	66.3%	61.8%	82.8%	*	93.2%	*	78.9%	61.1%
RHSP/DAP/FHSP-E/FHSP-DLA Graduates (Longitudinal Rate)											
Class of 2015	84.1%	82.3%	**79.6%**	74.1%	64.1%	85.4%	*	92.9%	*	93.3%	65.3%
SAT/ACT Results											
Tested											
Class of 2015	68.3%	76.5%	**69.4%**	69.4%	45.4%	74.1%	*	93.8%	*	83.3%	51.2%
Class of 2014	66.3%	76.7%	**69.3%**	68.8%	43.7%	74.4%	*	90.6%	*	82.1%	46.7%
Average SAT Score											
Class of 2015	1394	1695	**1642**	1442	1460	1681	*	1795	*	1743	1476
Class of 2014	1417	1693	**1644**	1430	1509	1653	*	1807	*	1706	1471
Average ACT Score											
Class of 2015	20.6	25.8	**25.1**	20.7	21.7	25.8	*	27.7	-	27.7	21.7
Class of 2014	20.6	25.8	**25.3**	20.5	23.3	25.4	*	28.0	-	25.7	22.1

For more information about this campus, please see the Texas Academic Performance Report at https://rptsvr1.tea.texas.gov/perfreport/tapr/2016/index.html.

Page 3

'?' Indicates that the data for this item were statistically improbable, or were reported outside a reasonable range.
'*' Indicates results are masked due to small numbers to protect student confidentiality.
'-' Indicates zero observations reported for this group.
'n/a' Indicates data reporting is not applicable for this group.

Figure 2. Sample page from 2015-16 New Mexico accountability report

College and Career Readiness (CCR)

High school students are expected to participate in at least one college or career readiness program. These activities include one of the following:

1) College entrance assessments (SAT, SAT Subject Test, PSAT, ACT, PLAN, Compass, ACT Aspire, or Accuplacer)
2) Evidence that the student can pass a college-level course (Advanced Placement, Dual Credit, or IB)
3) Eligibility for an industry recognized certification (Career Technical Education, SAM School Supplemental)

Points are given separately for students' participation and for their success. To be considered successful, students must meet established benchmarks. Details are in the School Grading Technical Guide on the PED website at http://webapp2.ped.state.nm.us/SchoolData/SchoolGrading.aspx.

CCR follows the Shared Accountability model used for cohort graduation rates. Cohorts are fully described in the Graduation Technical Manual on the PED website at http://ped.state.nm.us/Graduation/index.html.

50% or Higher
20% -50%
Below 20%

	All Students	Gender F	Gender M	Race / Ethnicity White	Afr Amer	Hisp	Asian	Am Indian	Economically Disadvantaged	Students with Disabilities	English Language Learners
Participation (% of Cohort)	28.1	23.6	30.7	3.9	<2.0	30.3	-	29.8	30.5	22.7	34.9
Participation (Pts)	1.41										
Success (% of Participants)	36.9	30.2	39.8	<2.0	-	36.4	-	55.3	33.3	10.9	28.5
Success (Pts)	3.69										
Percent of School's Cohort of 2015 Participating in Each CCR Opportunity											
ACT	<2.0	<2.0	2.3	<2.0	<2.0	<2.0	-	<2.0	<2.0	<2.0	2.8
PLAN	2.4	2.6	2.3	<2.0	<2.0	2.5	-	4.8	2.2	2.8	3.9
ASPIRE	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	-	<2.0	<2.0	<2.0	<2.0
SAT	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	-	<2.0	<2.0	<2.0	<2.0
PSAT	17.0	15.9	17.7	3.9	<2.0	18.5	-	13.3	18.6	13.5	23.4
AccuPlacer	3.7	<2.0	5.1	<2.0	<2.0	4.2	-	<2.0	4.4	7.3	4.0
Advanced Placement	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	-	<2.0	<2.0	<2.0	<2.0
Dual Credit	6.5	<2.0	9.5	<2.0	<2.0	6.5	-	16.5	7.3	<2.0	4.6
International Baccalaureate	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	-	<2.0	<2.0	<2.0	<2.0
Career Technical Education	2.6	<2.0	3.2	<2.0	<2.0	2.4	-	11.7	2.5	<2.0	<2.0
Compass	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	-	<2.0	<2.0	<2.0	<2.0
SAT Subject Test	<2.0	<2.0	<2.0	<2.0	<2.0	<2.0	-	<2.0	<2.0	<2.0	<2.0
SAM School Supplemental	2.1	3.9	<2.0	<2.0	<2.0	2.4	-	<2.0	<2.0	<2.0	2.4

Bonus Points

While most schools provide a sampling of athletics, club participation opportunities, and parent meetings, a few schools stand out among the rest. These schools are recognized for their extraordinary dedication to keeping students invested in school and their efforts in empowering parents to engage actively in their child's education.

- ☑ Student and Parent Engagement
- ☑ Truancy Improvement
- ☐ Extracurricular Activities
- ☑ Other

Participation

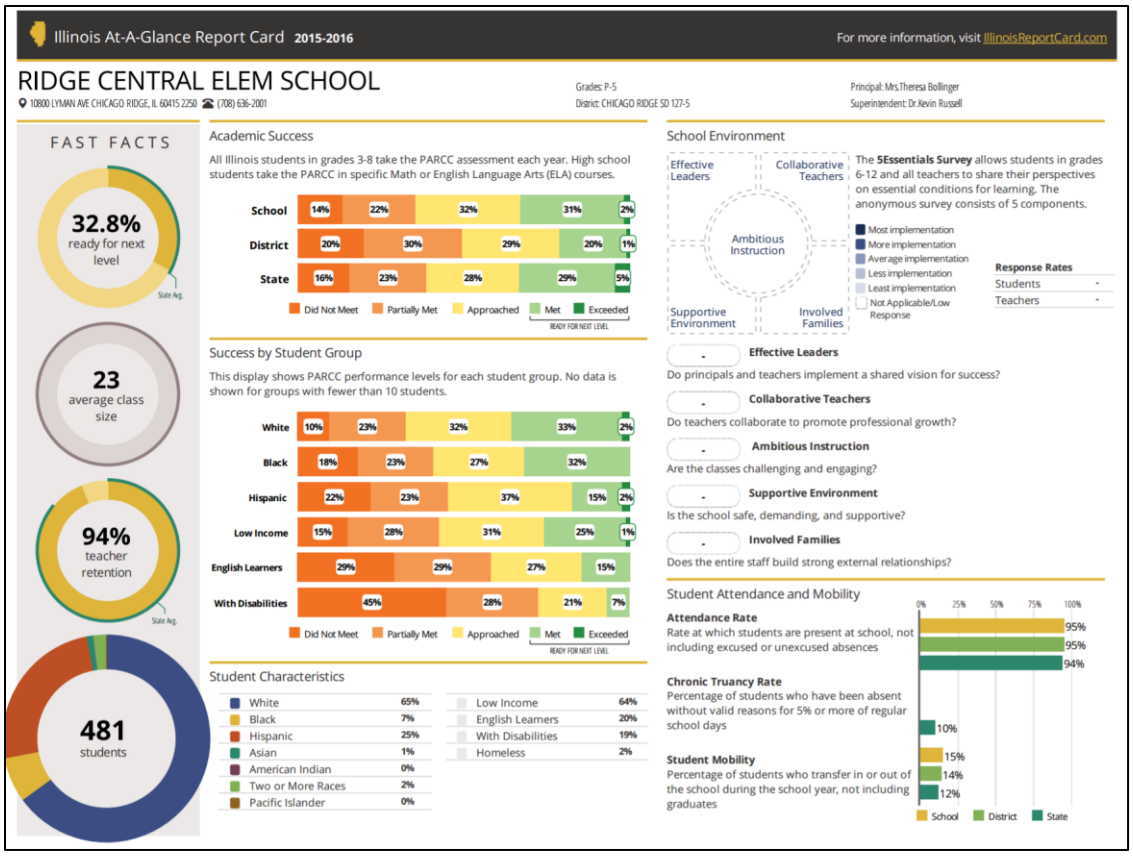
Schools must include all of their enrolled students in the annual statewide assessment. If the percentage of students is less than 95%, the school's letter grade is reduced by one grade. Supplemental Accountability Model (SAM) schools and small schools with fewer than 100 students receive special consideration.

Reading (%) 47

Math (%) 45

School exempted from penalty because of SAM status.

Figure 3. Sample page from 2015-16 Illinois accountability report



School accountability critics often argue over the content of accountability reports – over definitions of school quality and how to measure it – while curiously overlooking the form that this content takes (Kane & Staiger, 2001; Wiliam, 2010). Although this focus on content is not misguided, the inattention to form precludes researchers and practitioners from understanding the larger context in which school accountability reports participate. This article hopes to fill this gap by arguing for the importance of form in addition to that of content. With this in mind, throughout the article I will repeatedly make efforts to distinguish the role of both content and form. By deliberately separating the two, one can better examine the ways in which form mediates content.

To understand why form matters within this landscape of high stakes accountability, I will take several approaches. First, I consider the policy consequences embedded in these reports, both under current ESSA legislation and under the prior eras of NCLB and NCLB-waivers. From the passing of NCLB, to the Obama administration's NCLB waiver program, and through the signing of ESSA – how does the system of modern public school accountability work? What does the government require of states and, furthermore, what role do accountability reports play? What content must be included in these reports? What form should these reports take? What impacts – both intentional and unintentional – do these regulatory guidelines create?

Next, to better understand the impact of form on the reports themselves, I will draw on the field of technological theory, arguing that these reports are inherently political artifacts, constructed through a complex interaction of actors. This body of literature details the ways in which artifacts mediate our interactions with the world around us. Drawing on the work of Langdon Winner, this article examines the role that technologies play within society: How are specific design decisions made (or abandoned)? What impacts do these decisions have on those who interact with a specific tool or technology? Using this lens, there is no such thing as a neutral design – every decision prioritizes the needs of some actors over others.

Complimenting this theoretical approach, I then take a more practical and applied approach, drawing on existing research in the visual display of quantitative communication to see the ways in which form matters within data reporting. This literature suggests that even the smallest design decision necessarily incorporates bias. This is not to say that some designs are better or worse than others; but rather, that each design decision is a tradeoff, a choice to favor *this* presentation over *that* one, to support this reading of the information over all others. Although this literature is primarily from

outside the realm of public education, there is a small and emerging body of literature on data visualization within public education, and even within accountability reports themselves.

Together, these three avenues of approach – policy, technology, and data visualization – support a compelling view that form matters, that the visual display of information within accountability reports is not only as important as the content contained, but that these design choices fundamentally impact the very model of high stakes accountability in the United States.

Policy Impacts on School Accountability Reports

Holding schools accountable for providing a high-quality education is a broad and complex task. Federally mandated school report cards represent one small piece of the puzzle. In addition to federally-mandated reporting by states, there is often additional reporting by local school districts. For example, in my home town of Philadelphia, the School District of Philadelphia district has created its own accountability system for the more than 200 traditional public schools in the city. This system exists separate from the state of Pennsylvania's system, which means that every Philly public school receives both a state-sponsored accountability report and a district-sponsored accountability report, each offering a slightly different perspective on school quality. Moreover, school accountability reporting also exists outside of the government's purview. For instance, the website GreatSchools provides high-level rankings of school quality nationwide, primarily for use in real estate to help home buyers evaluate the quality of schools in their area. Finally, many of the most common mechanisms of school accountability are not tied to reporting at all, but rather emerge from one's own

experience with local schools, with friends' and family's opinions regarding which schools are good and which schools are not.

Among these many forms of accountability, federally mandated school accountability reports are unique in their universalism: a standardized report exists for every public school district and every public school in the United States. No other method of accountability reaches such a large audience in such a consistent way. Consequently, although these reports are by no means the sole mechanism for holding schools accountable, this article will narrowly focus on these reports due to their universal reach. Furthermore, in order to understand why these accountability reports are so central to the larger system of school accountability, it will benefit us to examine the origins, development, and implementation of federal accountability legislation.

NCLB and Establishing a Theory of Action for School Accountability

Broadly speaking, contemporary school accountability is inextricably tied to NCLB. NCLB emerged as an extension of President Johnson's Elementary and Secondary Education Act of 1965 (ESEA, H.R. 2362, 1965), legislation which first established Title 1 provisions, granting federal money to states with the expressed purpose of empowering states to exercise local control over the education of their students, particularly those from low-income families. Although NCLB retained much of the spirit of ESEA's reforms, the legislation represented a radical shift in the government's approach to implementation. Prior to NCLB, federal oversight of public education in the United States focused primarily on inputs – on the distribution and quality of teachers, on the specifics of curricula enacted in the classroom. Through NCLB, this focus on inputs was replaced with a focus on outcomes – onto measurable markers of student learning (Isaacs, 2003; Wong, 2008).

Put into practice, this shift towards outcomes became a shift towards assessment. NCLB enacted a sweeping student-level testing mandate for the nation's public school system. Under the law, states were required to implement summative assessments across multiple grades, in both math and English language arts; a concrete deadline was set for all students to reach proficiency on these assessments; and each school was expected to make clear strides towards this deadline in the form of adequate yearly progress (AYP) (NCLB, 2001).

In addition to this testing mandate, NCLB also instituted a clear system of incentives for schools and districts. If a school failed to meet AYP for two consecutive years, the district was required to provide students in that school with the opportunity to relocate to another school. If the same school failed to make AYP for another year, the school was also required to implement state-approved supplemental education services. Finally, after five consecutive years of failing to meet AYP, the school was required create a restructuring plan to change its leadership and governance structure, reopen as a public charter school, or turn over control to an outside entity or state education agency (NCLB, 2001).

Together, this notion of assessments tied to clear consequences created the environment of "high-stakes testing" that defines current K-12 accountability. Schools are responsible for their students' performance, specifically as measured by end-of-year assessment results. These results, in turn, have direct impact on school- and district-level autonomy. If students do not perform well on the test, the school loses control over funding and potentially management as well. Consequently, assessments become hugely important. Clear incentives are created for schools to do whatever it takes to improve student performance on the mandated tests – to the exclusion of other subject areas and academic outcomes (Ladd, 2001; Spillane, 2012).

Taking a step back, one can roughly sketch the broader theory of action behind this approach to school accountability as follows: All students deserve a high-quality education. The quality of education is measurable via students' performance on criterion-referenced assessments in key subject areas. If a school's autonomy is tied to student performance on these tests, that school will act to improve instruction and, therefore, the quality of education provided to its students.

Ultimately, this model of accountability depends on clear feedback (Isaacs, 2003). When striving towards any goal, it is helpful to know where we've been, where we're going, and most importantly, whether we're on the right track. In the world of K-12 education, this means knowing: How are our students performing now? How *should* they be performing at the end of the year? Is there any evidence of progress? As Kirby writes, "The goal of any program is to bring about desired outcomes. The goal of an evaluation of that program is to determine, through data analysis, whether the program did in fact have an effect on outcomes, and if so, the nature of the effect" (2002, p. 142). High stakes accountability is, by definition, an evaluation program. Its goal is to evaluate and assess the degree to which public education institutions have worked towards the benchmark established by federal legislation.

Importantly, this feedback loop does not – and was never intended to – stop at the walls of the school. Just as the legislation encouraged teachers and administrators to use data to drive decision-making *within* schools, NCLB expected that parents, families and the public at large will use data to drive decision-making *outside* of schools. As mentioned, the NCLB model established a (limited) system of school choice. Under the legislation, if a school repeatedly fails to meet AYP, parents must be given the choice to enroll their students at another, higher performing, public school (Shaul & Ganson, 2005). As a result of this choice option, parent and public perception matter. By failing

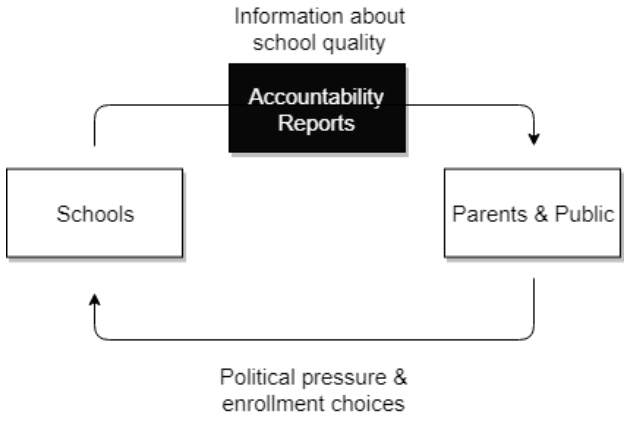
to prioritize standardized assessments, schools risk losing their customers. Consequently, one goal of NCLB "is to make sure that parents, particularly in disenfranchised neighborhoods, receive the necessary information on education options in a timely manner" (Wong, 2008, p. S178). By providing this information, parents become yet another lever pressuring schools to provide higher levels of service.

Perhaps the best summary of this theory of action comes from a speech given by President George W. Bush to the National Urban League:

> Accountability is an exercise in hope. When we raise academic standards, children raise their academic sights. When children are regularly tested, teachers know where and how to improve. When scores are known to parents, parents are empowered to push for change. When accountability for our schools is real, the results for our children are real. (2001)

And this is why accountability reporting matters. Accountability reports are the vehicle for feedback. Teachers know "where and how to improve" when they see student scores on standardized assessments – but where do they see these scores? Assessment data helps empower parents "to push for change" – but how do parents access this data? At the end of the day there is a single document, a single print or web report, that must do all of the heavy lifting. The entire system depends on feedback, and this feedback is – for better or worse – embedded into accountability reports. They are the cornerstone.

Figure 4. Role of accountability reports in school accountability



Ultimately, this is why accountability reports matter. These bureaucratic artifacts are integral to an incredibly complicated and politically fraught public discourse regarding the quality of education that our public schools provide to our nation. Yet, oddly, few have bothered to take a systemic approach to evaluating what they look like.

An Evolving Regulatory Landscape

Although this theory of action supporting public school accountability has remained relatively stable, the specifics of accountability legislation have been anything but. From the start of NCLB to present day, subsequent legislation has required states to publish an array of information on students and schools. As mentioned, NCLB was the first legislation to require states to develop and report AYP. By design, these AYP measures were incredibly prescriptive (Rose, 2004). Essential to NCLB was the idea that all states are responsible for ensuring that all students succeed. More specifically, NCLB set a goal for all states to ensure that all students were performing at or above grade-level within 12 years of the legislation's passing. AYP was designed to measure progress towards this twelve-year goal (Linn, Baker, & Betebenner, 2002). If students within a state were performing below expectations, the AYP goal would be set 1/12th of the way between current and expected performance. Further, to ensure that states focused their efforts on all students, particularly those in need, NCLB required AYP measures to be calculated for all students, as well as for several specific student groups (e.g., economically disadvantaged students, students with disabilities, etc.) (Linn et al., 2002). Notably, the legislation also dictates, quite specifically, what information must be publicly reported, including:

> (i) information, in the aggregate, on student achievement at each proficiency level ... (disaggregated by race, ethnicity, gender, disability status, migrant status, English proficiency, and status as economically disadvantaged ...)
> (iii) the percentage of students not tested (disaggregated by the same categories and subject to the same exception described in clause (i));
> (iv) the most recent 2-year trend in student achievement in each subject area, and for each grade level, for which assessments under this section are required;
> (v) aggregate information on any other indicators used by the State to determine the adequate yearly progress of students in achieving State academic achievement standards;
> (vi) graduation rates for secondary school students consistent with subsection (b)(2)(C)(vi);
> (vii) information on the performance of local educational agencies in the State regarding making adequate yearly progress, including the number and names of each school identified for school improvement under section 1116; and
> (viii) the professional qualifications of teachers in the State. (NCLB, 2001)

In contrast to this detailed description of what content must be included on accountability reports, guidelines on form are quite thin:

> (B) IMPLEMENTATION- The State report card shall be—
> (i) concise; and
> (ii) presented in an understandable and uniform format and, to the extent practicable, provided in a language that the parents can understand. (NCLB, 2001)

Beyond this, states are left to their own devices.

In many ways, this pattern of highly prescribed content and minimally prescribed form has held true in the shift from NCLB, to the Obama-era NCLB-waivers, and to the current ESSA regulations. The NCLB-waiver system was an attempt to free states from the stringent expectations of AYP. Though the notion of having all students performing at or above grade-level within 12 years may sound appealing, critics have derailed the policy as either naively optimistic or willfully ignorant of the realities of modern public education (Darling-Hammond, 2006; Forte, 2010; McMurrer & Yoshioka, 2013). With this in mind, the Department of Education allowed states to apply for waivers that would

grant exemption from the NCLB requirements in exchange for an alternate set of annual objectives created by the states themselves (McNeil & Klein, 2011).

By design, these waivers were less prescriptive in terms of content. The explicit motive for the waiver system was to empower states to create their own accountability measures, rather than forcing measures on them (McGuinn, 2016). However, the Department of Education did provide some guidelines. State submitted accountability plans were required to include clear college- and career-ready standards aligned to high-quality summative assessments. These standards would have to include both current performance as well as growth in performance over time, as well as a handful of other aligned indicators, including English proficiency rates for English language learners, measures of school climate and safety, and participation rates in alternate assessments ("State and Local Report Cards: Title I, Part A of the Elementary and Secondary Education Act of 1965, as Amended, Non-Regulatory Guidance," 2013).

Similarly, the waiver system provided no clear requirements on the form that reports must take. However, in a publicly disseminated guide to states, the Department of Education did provide guidelines for clear reporting:

> An effective report card presents student and school performance data in a manner that is clear, easy to understand, and accessible to all stakeholders and, most especially, parents of the students who are the focus of ESEA program… An effective report card is
>
> - Easy to read;
> - Accessible to the target audiences both physically and linguistically;
> - Accompanied by adequate interpretive information;
> - Supported by evidence that the indicators, other information, and suggested interpretations are valid; and
> - Coordinated across paper and electronic versions of report cards. ("State and Local Report Cards: Title I, Part A of the Elementary and Secondary Education Act of 1965, as Amended, Non-Regulatory Guidance," 2013).

As an extension of the waiver system, guidelines under ESSA remain largely the same. The legislation requires states to submit their own accountability plans, including

indicators of college- and career-readiness based on high-quality summative assessments, with results disaggregated by student groups. Additionally, the form of reports remains entirely up to the states themselves. As during the original NCLB legislation, ESSA requires reports to be "concise, presented in an understandable and uniform format, and accessible to the public." And, as during the waiver era, the Department of Education provides loose guidelines for report designers, providing them with high-level questions to guide their design process:

> - Does the report card design take into account feedback provided through the required parental consultation?
> - Does the report card reflect feedback based on different presentation formats presented to a variety of audiences representing likely consumers of report cards to ensure precise and clear communication of the data? If so, did the test audiences evaluate the use of font size, keys, graphs, page layout, instructions, and pagination?
> - Is the information on report cards in hard copy form? If so, is it shared online in the same format to ensure consistency across communication mediums?
> - Are the data available in both chart/graph and table format?
> - Do the graphics and artwork improve readability and maintain user interest?
> - Does the report card avoid using jargon not well known to parents?
> - Is social media embedded to allow parents to easily share information?
> - Do report cards include a brief narrative summary of relevant information for parents?
> - Does the report card link to historical information provided in previous years? (*Every Student Succeeds Act State and Local Report Cards Non-Regulatory Guidance*, 2017, p. 8,9).

Looking across these regulations and guidelines as they have evolved from NCLB, NCLB-waivers, and ESSA, there is a clear trend: accountability legislation provides states with little guidance and nearly no oversight over the form of accountability reports. In fact, the ESSA guidelines excerpted above represent the most direct and comprehensive feedback given to states, and yet even here the instructions are provided at an incredibly high level of abstraction. Moreover, while ESSA provides states with more flexibility in

defining the content and the form of their accountability reports, only the content is subject to regulatory review. While states must submit their ESSA plans (i.e. the data they intend to collect and report) for approval by the Department of Education, there is no mandated review of the reports themselves.

Altogether, ESSA makes a rather strong assumption that states will have the necessary expertise to represent accountability information in the most effective ways; however, even a cursory examination of the legislation and the legislative guidelines casts doubt on these assumptions. The legislation mandates that states create reports that are "concise, presented in an understandable and uniform format, and accessible to the public," without unpacking any of these terms. Who is "the public" referred to in this legislation; is it a monolithic group with homogenous interests or an amalgamation of various stakeholders? If the latter, should some audiences' needs be prioritized over others as designers work to create accessible reports? The guidelines also suggest including "graphics and artwork [that] improve readability and maintain user interest" but again, for whom? Readability is a loaded term, and one that depends highly on audience. Moreover, even if the audience was clearly defined, how might states achieve these goals? What types of design decisions improve readability? How might design decisions support concise and understandable data displays? What about the trade-off between concision and comprehensiveness (i.e., accurately including all mandated content without overwhelming audiences with minute detail)? The legislation leaves states with no actionable tools for achieving the vague goals it sets out for them.

Research in Data Visualization

To understand just how complex these goals are – how specific design decisions influence readability, understandability, and interest – one must turn to the existing

literature on the visual display of quantitative information. Since the late 18th century, scholars have spent considerable effort trying to understand how to best communicate quantitative information (Beniger & Robyn, 1978). In just the past fifty years, this effort has turned to into a well-documented area of research, with academics and practitioners validating industry best-practices in a more structured and deliberate way (Roberts & Gierl, 2009). Looking across existing work, this article traces common areas of practice and key design choices that go into contemporary data visualization.

In his foundational work, *The Visual Display of Quantitative Information*, Edward Tufte (2001) paints a broad picture of data visualization and its impacts. Detailing best practices of data presentation, Tufte argues that data visualization is no different from any other communicative act; "words, graphics, and tables," he writes, "are different mechanisms with but a single purpose—the presentation of information" (2001, p. 181). Just as one strives for clarity and precision when writing an essay, one must aim to be as clear and concise when communicating visually.

For Tufte (2001) this emphasis on communication – independent of medium – boils down to two key principles, ones that appear again and again in subsequent research. First, efficiency is key. For Tufte, all data visualizations exist for the express purpose of transferring information from author to audience. As a result, anything that does not serve this purpose is extraneous. Tufte uses the term "chart junk" to refer to such extraneous information – the colorful illustrations and decorative designs that accompany a chart or graph, but which do nothing to help convey the information therein. The goal, according to Tufte, should always be to minimize chart junk and to maximize the "data-ink ratio," the amount of ink on the page dedicated to displaying valuable information.

Second, Tufte argues for the importance of context. For Tufte (2006), visualizing data is not only about communication, but about rhetoric. A chart or graph not only provides information, it frequently provides an argument. For example, in the case of school accountability reports, the visual presentation of school-wide proficiency rates (theoretically) is making a suggestion about the quality of instruction at that school – whether it meets expectations, whether the quality has been changing over time, etc. Good arguments, Tufte argues, should provide context. Rather than willfully ignoring competing theories or shrugging off dissent, the most convincing arguments provide audiences with transparent information, allowing them to investigate and criticize until they are satisfied. In the same way, the best data visualizations are obligated to provide context, to avoid presenting half-truths and instead provide a complete and transparent view of the data (Tufte, 1990, 2001).

Echoing Tufte, Tukey (1990) argues that visual displays of quantitative information should be deliberate and purposeful. Like Tufte, Tukey details several baseline requirements for good charts and graphs – for example, they must be free of clutter and efficient in their use of ink and color. However, beyond this, Tukey argues that each chart and graph be chosen with clear rationale. In his words, "treating [a] visual display as a *tabula rasa* which will automatically and unbiasedly analyze the data, without need of computation, is to give up most of its value by asking it to do only what it does relatively poorly" (1990, p. 332). Different charts better serve different ends – whether comparing differences in a single measure across specific units, changes in a single unit over time, or the distribution of a measure across an entire population. Even within a specific visualization, Tukey argues, the specific choices of line thickness, line style, and visual emphasis should always serve the intended goal (1990, pp. 329, 330).

The Impact of Specific Design Elements

So, which visualizations serve which goals? MacDonald-Ross' (1977) pioneering – and still highly relevant – review of literature offers some initial insight. Like Tukey, MacDonald-Ross recognizes that there is no "best" visualization. In his words,

> No one graphic format is universally superior to all others… To choose the best format for a particular occasion one must decide: what kind of data is to be shown? What teaching point needs to be made? What will the learner do with the data? (1977, p. 401)

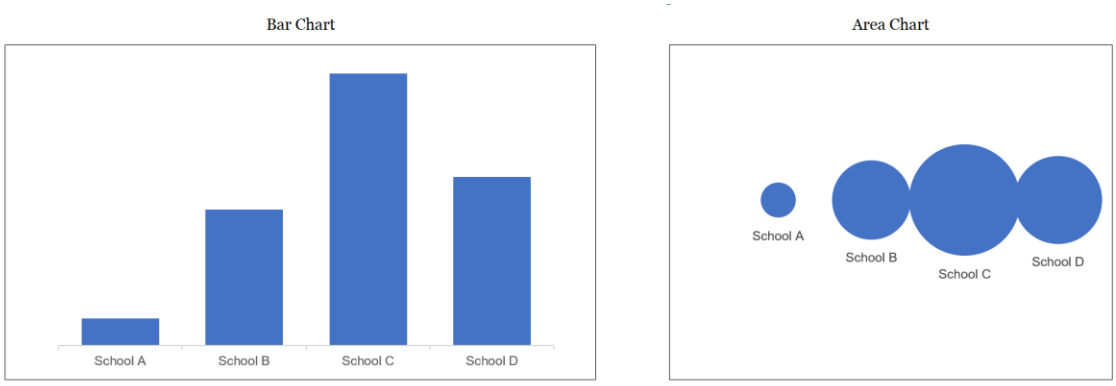
With this in mind, MacDonald-Ross moves through a litany of visualization techniques (e.g., bar charts, pie charts, tables, line graphs), detailing a brief history of research on each, and also offering some brief advice regarding which occasions, which data, and which teaching points are best served by each.

Interestingly, MacDonald-Ross considers the traditional numerical table to be one of the key methods of data visualization. Although the word "visualization" often conjures images of brightly colored bar charts and line graphs, a table is nonetheless a visual representation of data. Tables are particularly interesting as a visualization because they are so completely abstract. Unlike a bar or line, where users can perceive differences in values using length and position, numbers offer no clear indication of difference other than the symbols themselves (Macdonald-Ross, 1977).

Consequently, the choice between table and chart involves compromise. "The choice of table," MacDonald-Ross writes, "involves a complex trade-off between compactness, exactness, and ease of usage" (1977, p. 379). Because of their abstract nature, tables are difficult to decipher and to decode; differences in scores from one cell in the table to another can only be uncovered after significant mental effort. Bar graphs and line charts can significantly reduce this mental effort; however, tables are far more precise. Rather than measuring the height of a bar against labels on a vertical access, one can simply read an exact number of a page.

Moving from charts to graphics, Cleveland and McGill (1984) suggest that even across different graphical representations, certain design choices can help or hinder viewers' ability to accurately interpret the data. The authors presented participants with various graphical comparisons of two groups' performance, then assessed users' understanding of each. Based on the results, some graphical depictions allowed for easier and more accurate comparison than others – a finding which confirmed the arguments of several pioneers in this field of research (Brinton, 1914; Karsten, 1923). Participants made more accurate interpretations when the difference between groups was expressed on a common scale or shown via length (e.g., bar charts). Participants made less accurate interpretations when differences between groups were expressed as area, volume, or difference in color saturation and hue. To see why this is the case look at Figure 5, below. In this figure, hypothetical graduation rates are displayed for several sample high schools, with the exact same values displayed as a bar chart and an area chart. In this example, the graduation rate at school C is exactly double that of School B. Looking at the bar chart, one can see this relationship rather clearly. Even in the absence of data labels and axes, the bar for School B appears about half the size of School C. However, in the area chart, this relationship is much harder to spot. Here the differences between School B, C, and D look less dramatic, as the eye tends to notice the difference in diameter of each circle, rather than the difference in surface area. This same complication arises with pie charts (a subset of area charts) as the eye must compare the surface area of each slice of the pie relative to one another, with each at different degrees of orientation.

Figure 5. Comparison of bar and area charts



Similarly, Shah, Mayer, and Hegarty (1999) suggest that chart makers can significantly influence how audiences will interpret data by choosing how to "chunk" or group elements together. The authors argue that "it is not enough that graphs are merely technically correct in presenting relevant information." Instead, the format of the graph – the ordering of elements, the visual organization of the page, and other aesthetic choices – is what truly matters. To demonstrate this point, the authors asked participants to describe several charts that depicted US Civil War population size and distribution (rural vs. urban) across the North and South. When this information was depicted in a line graph, rather than a bar graph, participants were more likely to describe trends over time (from year-to-year) rather than trends across groups (North vs. South). In other words, different graphs, depicting the same information, can prime audiences to favor one interpretation over another. However, the authors found that the choice of graph was actually less important than the grouping of information within a graph. When the authors "chunked" information by year, users tended to describe year-to-year patterns in the data *regardless of if the data was presented as a line or a bar*. When lines and bars chunked information by group (North vs. South), audiences described changes over time (Shah et al., 1999).

In addition, several researchers have noted that type of chart or graph can impact the accuracy of readers' interpretation. In other words, some charts are better than others when it comes to understanding data. For example, Schonlau and Peters (2012) randomly assigned study participants with one of several data displays: tables, bar charts, and pie charts. Each of these charts displayed the exact same numbers; only the presentation varied. However, when asked to answer a series of questions about the data presented, participants who viewed the table and the bar charts significantly outperformed those who viewed a pie chart. In particular, for estimation of absolute proportions, for judging whether two categories are equal, and for estimating differences in size the authors found that tables and bar charts were more likely to lead participants to the correct answer (echoing Cleveland and McGill's findings described above with respect to comparisons using bar charts and area carts).

Similarly, Stewart, Cipolla, and Best (2009) studied the impact of extraneous information within data displays on participants ability to accurately answer specific questions. The authors' hypothesis was that extraneous information – Tufte's "chart junk" – would negatively impact participants' interpretation of data. To test this, the participants were randomly assigned to either a 2-dimensional (2D) or 3-dimensional (3D) chart display. Importantly, the 3D display did not show a third dimension of *data,* but instead added a visual shadow on top of the existing 2D display. Despite both treatment groups seeing the exact same information, participants in the 2D group performed significantly better than those in the 3D group on a series of questions about the data. In particular, the authors found a statistically significant interaction between chart type and question difficulty. The harder the question, the greater the difference between the two groups. Put another way, the harder the question, the more distracting the "chart junk" became.

These last few studies deserve deliberate consideration. Both studies suggest that *some data displays are simply better than others in conveying certain types of information*. The choice of presenting data as a table or bar chart or pie chart or line graph will not only influence how audiences access information, it will influence how audiences understand that information. Thinking back to the world of high stakes accountability, the theory of action behind these reports is that transparency will precipitate change. By shining a light on school performance, these reports empower stakeholders to pressure decision-makers into improving the quality of service. But, these studies suggest that transparency is more than publicity. It is equally related to the mode of presentation. Certain presentations encourage misinformation.

Beyond the choice of visualization, there is also significant research on finer grained details. Take, for instance, the use of capitalization. Report designers occasionally rely on capital letters in headings and throughout the reports to call attention to specific text. However, as far back as the 1940s, Breland and Breland (1944) found that readers were nearly 20% slower reading text in all capital letters than text with traditional capitalization. In this same vein, Coles and Foster (1977) found that bold text was more effective in drawing readers' attention than text in all capital letters. Even details so minute as the length of each line of text on a page, to the spacing between those lines of text has been shown to impact readers' comprehension of material (Katzir, Hershko, & Halamish, 2013).

In this same vein, several researchers outline the importance of deliberate and judicious use of color to emphasize key points. Winn (1991) suggests that color is effective precisely because it does not require cognitive effort –audiences are able to discriminate color "pre-attentively", differentiating elements and recognizing patterns via color much more quickly than via more cognitively challenging means (e.g.,

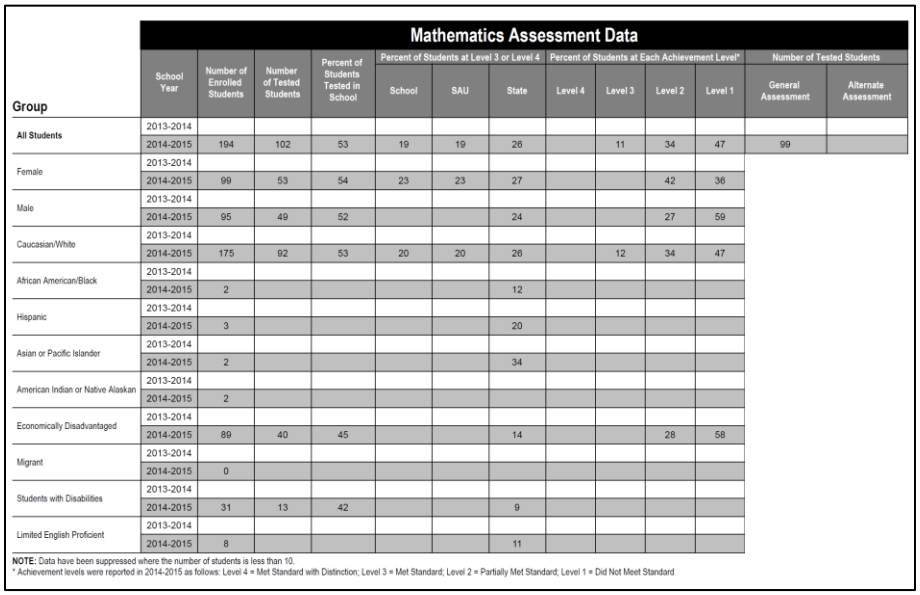
evaluating the relative size of values in a data table). In support of this claim, Benbasat and Dexter examined how the presence or absence of color on data displays influenced participants' perception of the reports' accuracy, as well as participants' ability to actually make accurate judgments based on the information itself. The researchers found that the use of color led participants to hold the report itself in higher regard; however, color only improved participants' comprehension and decision-making when researchers put participants under a time constraint (Benbasat & Dexter, 1986, p. 77). Complicating these findings even further, Vaiana and McGlynn caution report designers to keep in mind that nearly 10% of all readers experience some form of color-blindness, greatly impacting the effect of color on their perception and comprehension of data reports (Vaiana & McGlynn, 2002, p. 7).

More broadly, Gribbons (1992) describes how any array of choices, including type, color, and spacing, but also horizontal and vertical alignment, can cue the reader towards specific pieces of information. Gribbons lays out several best practices for creating clear and coherent reports. Among other things, Gribbons argues that information design authors work towards the principles of *selective employment* and *consistency* (1992). Selective employment refers to just that: the selective and deliberate use of design elements throughout a report. Imagine a tiny splash of red paint on a blank white canvas; the selective employment of color calls immediate attention to the brush strokes. Now imagine a canvas covered top to bottom in the same red paint. Suddenly, that first brushstroke is indistinguishable. Similarly, when a data report includes selective use of design elements – for example large bold fonts or deliberate limited highlights – those elements will stand out and guide the audience to key information. If, on the other hand, the entire document is riddled with yellow highlights and bold italicized text, that emphasis is lost. Readers aren't guided, but rather frustrated.

Coming from a cognitive science background, Vaiana and McGlynn (2002) echo Gribbons' – point: consistency in design elements supports readers' accurate interpretation of data-based reports. Drawing on a wealth of research in cognitive science, Vaiana and McGlynn argue that when section headers are clearly distinguished with the same typography, when tables are formatted consistently, and when charts share the same orientation, colors, and labels, readers can more easily navigate information. In their words, "A document's structure significantly affects how well readers understand and remember the information it contains" (Vaiana & McGlynn, 2002, p. 5). Without consistency in these elements, every new design element is yet another roadblock in the readers' path slowing their progress and potentially obfuscating accurate interpretation of the text.

To see these points in practice, consider Figures 6 and 7. Figure 6 is an excerpt from Maine's 2015-16 state reports showing assessment performance by student group. The Maine report demonstrates consistent use of colors, fonts, white space, and data displays throughout. Although some readers may be overwhelmed by the full-page data table, the table itself is consistent, with alternating white and gray shading, equal spacing between lines, and consistent font sizes for row labels, column headers, and data points themselves. Figure 7 is an excerpt from New Jersey's 2014-15 state reports, showing demographic information for a local public high school. In contrast, the New Jersey report has almost no consistent design principle. There is no deliberate or reliable pattern to the placement of data or the use of white space – each chart and table seems shoehorned into a cluttered page. There are five different types of data displays used (bar chart, line chart, stacked bar chart, pie chart, table), with little consistency in design elements and font sizes across each. As a result, it is much harder to locate information or even compare information on the New Jersey excerpt than with the Maine excerpt.

Figure 6. Example of typography, layout, and white space in Maine

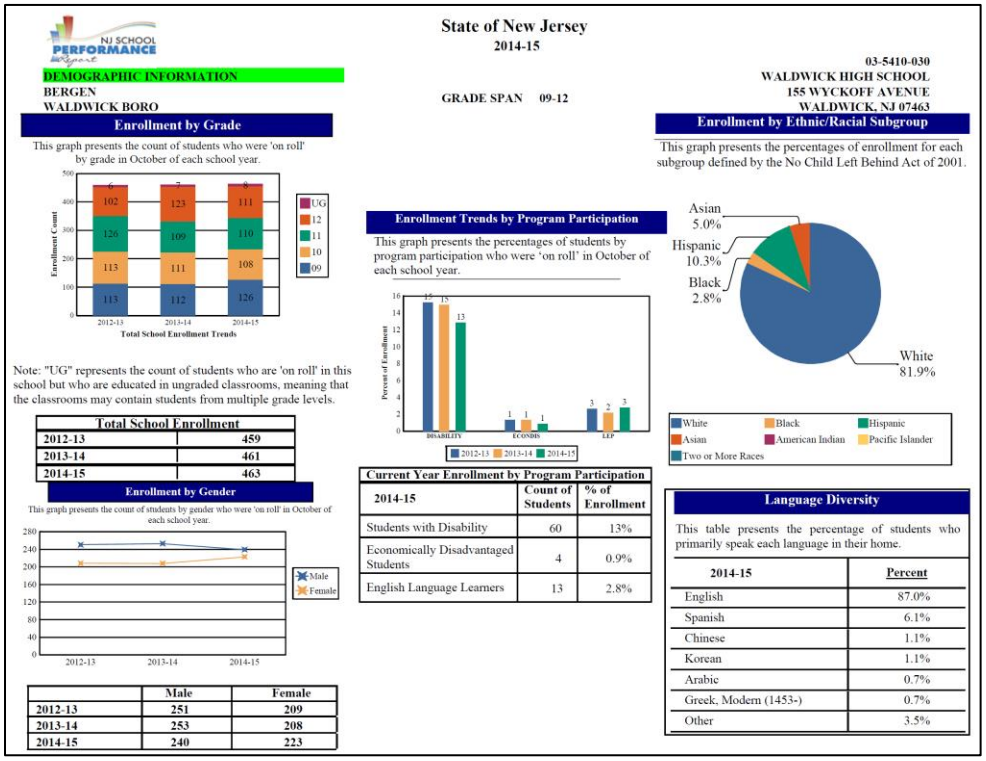


Mathematics Assessment Data

Group	School Year	Number of Enrolled Students	Number of Tested Students	Percent of Students Tested in School	Percent of Students at Level 3 or Level 4: School	SAU	State	Percent of Students at Each Achievement Level*: Level 4	Level 3	Level 2	Level 1	Number of Tested Students: General Assessment	Alternate Assessment
All Students	2013-2014												
	2014-2015	194	102	53	19	19	26		11	34	47	99	
Female	2013-2014												
	2014-2015	99	53	54	23	23	27			42	36		
Male	2013-2014												
	2014-2015	95	49	52			24			27	59		
Caucasian/White	2013-2014												
	2014-2015	175	92	53	20	20	26		12	34	47		
African American/Black	2013-2014												
	2014-2015	2					12						
Hispanic	2013-2014												
	2014-2015	3					20						
Asian or Pacific Islander	2013-2014												
	2014-2015	2					34						
American Indian or Native Alaskan	2013-2014												
	2014-2015	2											
Economically Disadvantaged	2013-2014												
	2014-2015	89	40	45			14			28	58		
Migrant	2013-2014												
	2014-2015	0											
Students with Disabilities	2013-2014												
	2014-2015	31	13	42			9						
Limited English Proficient	2013-2014												
	2014-2015	8					11						

NOTE: Data have been suppressed where the number of students is less than 10.
* Achievement levels were reported in 2014-2015 as follows: Level 4 = Met Standard with Distinction; Level 3 = Met Standard; Level 2 = Partially Met Standard; Level 1 = Did Not Meet Standard

Figure 7. Example of typography, layout, and white space in New Jersey



NJ SCHOOL PERFORMANCE Report

DEMOGRAPHIC INFORMATION
BERGEN
WALDWICK BORO

State of New Jersey
2014-15

GRADE SPAN 09-12

03-5410-030
WALDWICK HIGH SCHOOL
155 WYCKOFF AVENUE
WALDWICK, NJ 07463

Enrollment by Grade

This graph presents the count of students who were 'on roll' by grade in October of each school year.

Total School Enrollment Trends

Note: "UG" represents the count of students who are 'on roll' in this school but who are educated in ungraded classrooms, meaning that the classrooms may contain students from multiple grade levels.

Total School Enrollment	
2012-13	459
2013-14	461
2014-15	463

Enrollment by Gender

This graph presents the count of students by gender who were 'on roll' in October of each school year.

	Male	Female
2012-13	251	209
2013-14	253	208
2014-15	240	223

Enrollment Trends by Program Participation

This graph presents the percentages of students by program participation who were 'on roll' in October of each school year.

Current Year Enrollment by Program Participation		
2014-15	Count of Students	% of Enrollment
Students with Disability	60	13%
Economically Disadvantaged Students	4	0.9%
English Language Learners	13	2.8%

Enrollment by Ethnic/Racial Subgroup

This graph presents the percentages of enrollment for each subgroup defined by the No Child Left Behind Act of 2001.

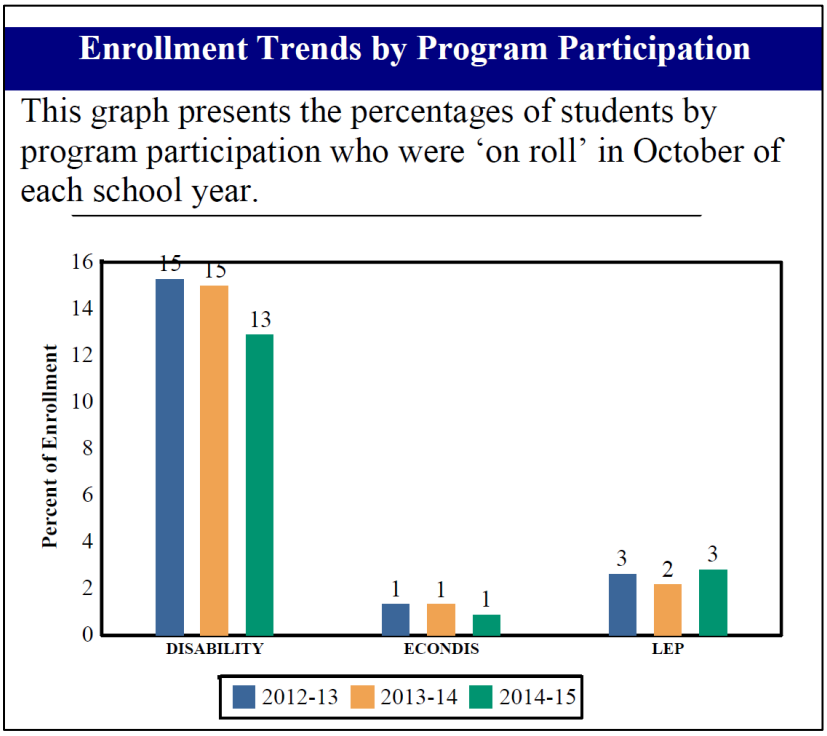
Language Diversity

This table presents the percentage of students who primarily speak each language in their home.

2014-15	Percent
English	87.0%
Spanish	6.1%
Chinese	1.1%
Korean	1.1%
Arabic	0.7%
Greek, Modern (1453-)	0.7%
Other	3.5%

Finally, moving from macro principles to more micro ones, practitioners make a strong case for paying attention to fine-grained details of charts and graphs. “The basic elements,” writes Cleveland, “tick marks, scales, captions, plotting symbols, reference lines, keys, and labels… are critical controlling factors whose proper use can greatly increase the accuracy of the information that we visually decode from displays of data” (1994, p. 16). Tufte (1997) emphasizes the importance of tightly linking visual and textual elements; without clear labels and markers, charts become cumbersome to decode. Clearly depicted axes titles and chart labels help remove abstraction and improve legibility. In addition, Huff (1993) emphasizes the importance of graphical scale. When authors choose to truncate or otherwise manipulate the ordinate and abscissa (e.g., by plotting a chart from 80 to 100, rather than 0 to 100), the resulting image is misleading. By truncating axes, small changes in absolute values appear prominently; by stretching axes, large changes in absolute values are diminished. Figure 8 provides a mixed-bag in terms of applying these lessons. The chart shows the percentage of students at the school who have a disability (“DISABILITY”), who are economically disadvantaged (“ECONDIS”), and who have limited English proficiency (“LEP”) over a three-year span. The bars are visually linked to the legend by color and there is a common graphical scale on the y-axis for all groups and all years; however, while the chart includes labels for each data point, the labels themselves are abbreviations used without explanation.

Figure 8. Excerpts from New Jersey's accountability report with "chunking" by student group



Looking across this research, one begins to see the enormity of the task set before states as they embark on accountability report design and the inadequacy of the ESSA reporting guidelines purported to assist states in achieving this goal. In order to make concise, understandable, and interesting reports, report designers must have a clear understanding of the content they are reporting and the audiences they are serving, as well as a profound mastery of data visualization. They must understand how to choose the right representation for the right content, balancing the tradeoffs each representation entails, as well the technical expertise to bolster those representations with rational design elements to meet the specific needs of their specific audiences.

Visualizing Education Data

With this in mind, it is worth acknowledging that several researchers have already attempted to transfer these generic lessons of data visualization to the specifics of K-12 public education. While their conclusions closely mimic the theorists' considered above, this body of work provides a helpful frame for our own discussion of accountability reports. For example, in their "Guidelines for Effective Score Reporting", Aschbacher & Herman (1991) attempt to provide a comprehensive view of assessment reporting practices, along with guidelines for creating more effective assessment reports. In doing so, the authors lay out several key recommendations:

1. Know the audience and purpose
2. Keep it simple
3. Be clear, accurate, comprehensive, and balanced
4. Use techniques to capture and focus the reader's attention

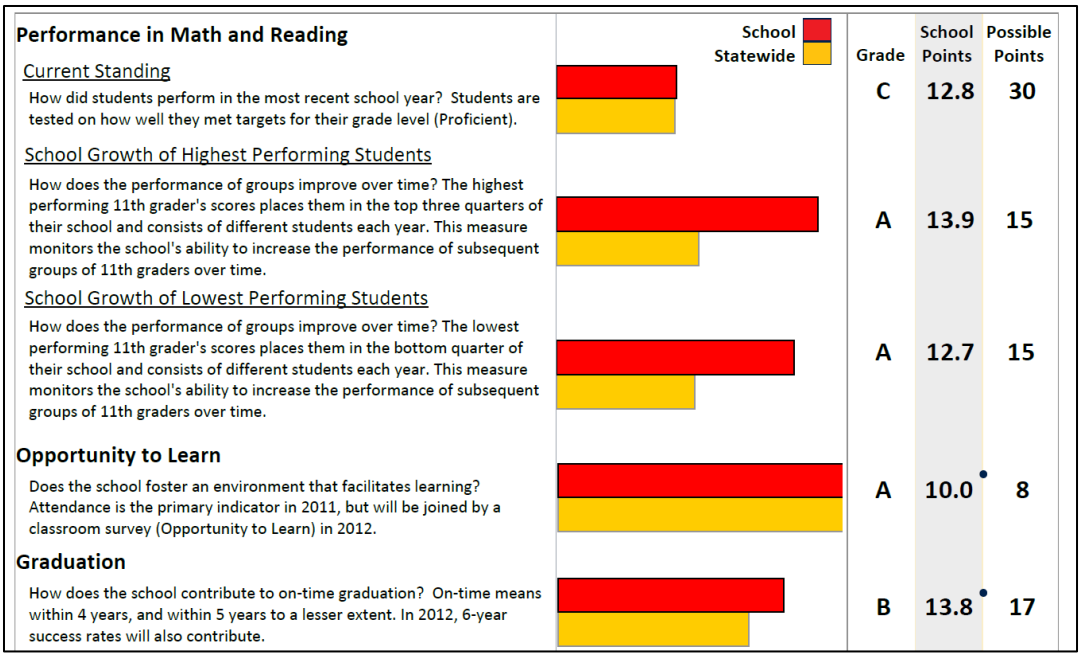
Recommendations three and four echo the research already reviewed. Whether one is dealing with agricultural, financial, or education data, there is value in "chunking" as a technique to focus attention and in choosing the right visualization (table vs. graphic) to enhance clarity and accuracy. The first two recommendations, however, offer something new. In particular, they suggest that one's choice of visualization should reflect both the type of data and the type of audience involved.

Within education contexts, this relationship between type of data and type of audience is complex (Hattie, 2009). Education data revolves around the student – for example, each student's individual outcomes on standardized assessments. This student-level information is instrumental to the audience of teachers in the classroom. Student-level scores allow teachers to review individual students' strengths and weaknesses, and to respond immediately to help improve student performance (Boston, 2002). For

accountability purposes, however, student-level information is less valuable. While individual student results provide information about a single student's performance, accountability systems are interested in school-level performance. When creating accountability measures, individual student scores are aggregated at the group and/or school level (e.g., percent of economically disadvantaged students scoring proficient or advanced, percent of all students scoring proficient or advanced). These group- and school-level measurements are then distributed widely to both administrators and parents (Kane & Staiger, 2002).

With this school-level view in mind, researchers offer several guidelines for reporting academic data to the public. Goodman & Hambleton (2004) remind report designers that public audiences often have little experience interpreting highly technical score reports. Consequently, they advocate that report designers include explanatory text and/or supplementary information to walk readers through the appropriate interpretation of the report results. Figure 9 provides a good example of this type of explanatory text. Each metric includes a brief non-technical description of what is measured, helping orient readers to the data displayed at right. Similarly, they encourage authors to present more general information before moving to more specific information. Rather than overwhelm lay audiences with technical detail, it is better to start with the most familiar and general concepts, then build on those concepts when presenting more nuanced information (Goodman & Hambleton, 2004; Hattie, 2009).

Figure 9. Example of explanatory text from New Mexico's accountability report



In this same vein, the "keep it simple" maxim is particularly important in education settings. Hambleton & Slater (1996) conducted detailed interviews with policymakers and educators, asking their participants to interpret the results of the National Assessment of Educational Progress (NAEP). In general, participants struggled with both the complex jargon and the technical statistics within the reports. Additionally readers of the report also struggled with mathematical inequality symbols, references to "statistically significant" values and "standard errors", as well as cumulative score descriptions (e.g., "X% of students were at *or above* level Y."), Similarly, Zwick et al. (2014) find that practitioners often struggle with statistical concepts like standard error, and generally prefer "information presented as short, easy-to-read pieces." Reports with large blocks of dense text are often misinterpreted, if not completely ignored, by most readers.

Additionally, research specific to education contexts echoes the finding discussed above: when audiences are provided with exactly the same content, but in slightly different formats, audiences often walk away with different understanding of that content. Research by Hastings, Weinstein, and Van Weelden (2008) illustrates this point quite well. Focusing specifically on school choice, rather than accountability more broadly, the authors used the introduction of school choice in the Charlotte-Mecklenburg School District to evaluate how the information (i.e., school reports) impacted parents' and students' choices of which school to attend. In the experiment, all families in the district were given access to broad information about school quality, including a choice handbook with demographic information about the school and school-provided qualitative information, as well as online school profiles, which included standardized test scores, suspension rates, and attendance rates, amongst other data. However, families in certain regions were given access to additional "simplified information sheets … specialized for each child" (2008, p. 13). These simplified reports included a list of schools available to that student, along with either (A) a single data point for each school representing students' average test score performance at that school, or (B) this same single score for test score performance printed alongside the students' odds of being accepted into that school. Comparing treatment groups, the authors found that these simplified reports led families, on average, to show an increased preference for schools with higher average performance than those families who did not receive the reports. In other words, assuming better test scores imply a better education (an assertion which itself is highly problematic), one could conclude that better information leads to better choices.

Although this particular experiment complicates the notion of school accountability with the intricacies of choice and subjective definitions of quality, the

results point to a more fundamental issue: *the presentation of data matters*. Hastings, Weinstein, and Van Weelden suggest that the way authors present data will influence the ways in which audiences respond. The authors describe their supplemental reports as "lowering information costs". In other words, the existing data displays – whether in the choice handbook or in the online school profiles – were costly displays. It requires a lot of time and effort to locate, assess, and compare the data provided. By contrast, the simplified and personalized reports were low-cost displays. Families could much more quickly and easily compare tests scores and odds of entry – and, as a result, families with access to these low-cost displays were more likely to prioritize test scores in their decisions.

The key takeaway here is not that test scores are the best indicator of a good school, or that simplified displays are better than more complicated displays; the takeaway is that *the display itself influences behavior*. Different designs encourage readers to take different actions. The current model of high stakes accountability suggests that the public will use accountability reports to pressure schools and districts into improving their services; however, these findings suggest that different report designs will lead readers to make different decisions.

Contemporary Accountability Reports

Despite the growing body of research on education-specific reporting practices, research on federally mandated school reporting practices is quite limited. In examining past research, I found only two academic publications that take up the topic in detail: a research brief by Education Commission of the States (ECS) (Mikulecky & Christie, 2014) and an academic article investigating the influence of score format on users' interpretations of school quality (Jacobsen, Snyder, & Saultz, 2014).

In their research brief, Mikulecky & Christie (2014) set out to answer three concrete questions regarding current school accountability reports: (1) Are the report cards easy to find? (2) Are the report cards easy to understand? (3) What are the current best practices for measuring performance? To answer each of these three questions, the authors compiled a 50-state database of school accountability reports and then put those reports in front of three distinct audiences.

First, the authors asked practicing researchers to assess the accessibility of the reports – in other words, how hard is it to sit down at a computer and actually find the documents? More specifically, the authors enlisted three existing ECS researchers to review each of the school accountability reports in the database, rating them "from 1 (unsatisfactory) to 3 (excellent) in the following categories: Findable, Readable, Understandable and Graphics." The authors suggest that for the "graphics" category, the rating was a direct response to the question "Were graphics used well to convey the information?"

Ultimately, the report suggests that these reports are not easy to find. In their words "even those experienced in online research had difficulty" locating the publicly accessible documents (Mikulecky & Christie, 2014, p. 6). Unfortunately, little additional detail is given on the researchers' findings regarding whether the reports have "understandable graphics." The only information included are excerpts from unnamed participants' responses to the Arizona, Illinois, and Ohio reports, which are very general reactions (e.g., "The graphics are well-done and convey information at a glance", "I also liked how the graphics were interactive") along with the reviewers' "Likes" and "Dislikes". These, too, are presented as direct quotes from participants with little context. For example, the "likes" include: "The graph titles also provide graphical information by hovering over the text," and, "I really like the overview on the first page with the

snapshot and basic graphs." In contrasts, the "dislikes" include: "There are a bunch of nice charts and graphs, but you have to click on each thing separately to see them," and, "Nearly unreadable. It was very difficult to understand what was being tracked or scored" (Mikulecky & Christie, 2014, p. 7).

Second, the authors assembled a panel of experts to review the metrics included in each of the reports and to offer their recommendation on which metrics prove most essential to high-quality accountability reporting. Based on conversations from this twelve-member panel (named the ECS School Accountability Advisory Group) the authors provide five "essential indicators" that every state should include in their accountability reporting: student achievement, student academic growth, achievement gap closure, graduation rates, and college and career readiness.

Finally, the authors asked parents to review individual reports and provide feedback on what makes a report easy or difficult to understand. Specifically, Mikulecky and Christie recruited 14 parents, representing "a mix of educational attainment, ethnicity, income levels, and geography" with students from kindergarten age through high school age. The authors asked each participant to review all 50 reports, rating each from 1 (unacceptable) to 5 (excellent) on three categories: "easy to read", "provides sufficient data", and "useful".

For present purposes, this last research question is most pertinent. Although accessibility and best-practices are important, the question of interpretation and understanding speaks most directly to our current topic. Unfortunately, the policy brief is light on detail. No statistical summary information is provided from the 700 report card reviews conducted by the 14-parent panel. Instead, the authors only report general reflections based on their own interpretation of the data. That said, according to the authors, the parents in this study echoed previous research on the visualization of

education data. Broadly, parents prioritized the clear and concise presentation of information, while complaining of information overload and complex mathematical calculations. In those reports rated most understandable by parents, parents applauded the use of clear and colorful graphics, as well as the inclusion of easily accessible contextual information on specific measures and data points. Conversely, in those reports rated least understandable, participants expressed frustration at "data in tables not clearly labeled or explained," as well as with the glut of information provided. One participant described a particularly lengthy report as akin to "reading a corporate financial report of 20 pages to get information." Although interesting, the Mikulecky and Christie paper is somewhat light on empirical evidence.

In contrast, Jacobsen, Snyder, and Sualtz (2014) bring a significantly stronger methodology to an even more precise aspect of school accountability reports: the overall score. In their research, the authors similarly compiled reports from all 50 states, as well as the District of Columbia, and eight large urban cities. Looking across this database, they define four distinct ways in which school accountability reports display their overall results: (1) performance indices, (2) letter grades, (3) performance rankings, and (4) percent of students attaining a defined goal.

Within the context of this research, a *performance index* refers to a single overall score calculated based on a schools' performance on any number of data points. For example, with a student receiving a 3.2 overall GPA based on their grades in multiple courses throughout the year, the 3.2 is an index that represents multiple assignments, assessments, etc. In contrast, *letter grades* refer to the familiar A, B, C, D, F system, where A represents the top level of performance and F the bottom. In this framework, schools are grouped into large buckets of performance differentiating high and lower performers. *Performance rankings* are similar to the letter grade scale; however, in this

system the familiar A, B, C system is replaced with descriptive text (e.g., Advanced, Proficient, Basic, Below Basic). Finally, *the percent of students attaining a defined goal* is just that: a single number that reports on the percent of a school's population that has met a predefined benchmark (e.g., percent of students meeting or exceeding expectations on end-of-year assessments). With these four formats in mind, the authors ask whether the format of overall results (e.g., performance index vs. letter grade) impacts how audiences interpret a school's overall quality.

To get at this question, the authors created three fictional schools: one low-achieving, one high-achieving, and one achieving somewhere in between. Next, the authors assigned ratings to the school's performance across each of the four formats. So, for example, the high-achieving school would receive a 95 out of 100 on an index scale, an A on the letter grade scale, and an Advanced on the performance ranking scale. The quality of the fictional school remained the same, but the description of that quality differed.

Once these rankings were complete, the authors asked a nationally representative panel of 1,111 respondents to evaluate the quality of a random school, based on a randomly assigned format. When reviewing a school, participants were asked one of three interrelated questions, each scored on a 7-point scale: 1. How satisfied are you with the school's overall performance? 2. How well does the school meet your expectations? 3. How close is this school to your ideal school? Results were then compared.

Overall, the format of results significantly influenced the panelists' interpretations of school quality. Unsurprisingly, participants in each experimental condition rated the high-performing school well and the low-performing school poorly. However, based on ANOVA results, participants who viewed performance as letter grades rated the high-performing school more highly than those who saw performance

depicted in any other way ($p < .05$). Similarly, participants who viewed letter grades tended to rate the low-performing school even *lower* than the others rated the same school using the other approaches ($p < .05$). For the high-performing schools, the difference was just over half a point on the 7-point scale; for the low-performing schools, the difference was just over one quarter of a point.

Although seemingly small, these differences suggest that the choice of representation matters. Without ever improving the actual quality of their schools, a school district could change public perception simply by changing the method of reporting. Switch from a 0-100 index to an A-F scale and suddenly your high achieving schools look even better. Switch from an A-F scale to performance ranks and your low achieving schools look slightly better off. By changing the representation *of the very same data* authors can affect audiences' interpretations.

This finding – that representation influences interpretation – emphasizes the importance of better understanding the current state of school accountability reports. School accountability is an enormous political machine that influences the lives, and future livelihood, of millions of American families. At the heart of this machine is the belief that communication will drive change. Provide administrators and parents with clear information about school quality, and those administrators and parents will push for reform. But, as the literature in data visualization shows, "clear" information is often elusive. Viewed from a technological perspective, accountability reports emerge as deliberately designed documents, as the end result of multiple choices regarding what to include, how to include it, and what form it will take on the page. Existing literature suggests that these different forms of representation may help or hinder a reader as they work to make sense of the underlying information. Here, it appears, they may actively shape a reader's understanding as well.

The Need for Future Research

Together, these findings make a strong case for additional research into the design, development, implementation, and interpretation of contemporary accountability reporting. The broader research on data visualization practices, as well as the specific research on data visualization in education, suggest that representation matters. Different representations may help or hinder different responses to the information itself. Representation is inherently political.

This fact complicates the role of accountability reports in the broader sphere of high-stakes school accountability. School accountability depends on the transparent and accurate communication of information to the public; legislation requires that states inform the public on the state of their schools. It is up to the public to use this information, to evaluate whether their schools are supporting their communities, and, if not, to pressure them to change. But this model naively assumes that states will provide information in an unbiased way – that with perfect information, the consumers of public education will make perfectly efficient choices, driving demand and influencing supply.

The research on data visualization suggests this model is flawed from the start. There is no unbiased way to present information; "there is no such thing as 'facts displayed' pure and simple." One might represent the exact same data about school quality in multiple ways; regardless, each represents a different form for the exact same content. And, more importantly, each may lead audiences to take different action – for example, to praise the quality of their school, or perhaps to push for immediate reform.

While this all may seem abstract, the Mikulecky and Christie (2014) research helps put a finer point on things. In their research on summary school ratings, the exact same information represented in slightly different ways led participants to judge schools

more or less harshly. Though one may justifiably argue the size of the effect, the bigger issue here is: 1) prior research suggests there is good reason to believe there is an effect, 2) this effect may have direct impact on the validity of our public school accountability system, and 3) no one has yet investigated this phenomenon.

This alone is a strong defense for additional research into nearly every aspect of school accountability reports. First, what is the status quo of accountability report design? How does the presentation of information compare from state to state and within state? What is the range and variation of data visualization techniques and of data elements across the nation, even when controlling for content across states or within states over time? Do the visualization choices in some states prioritize one interpretation of the data over another? Has this prioritization changed over time?

Second, how are these decisions made? The design of a document necessarily mediates audiences' interpretation of the message; yet, to what extent is this mediation deliberate? In other words, do the authors of these reports have an agenda? If so, do they consider the impact of minute but meaningful design choices on that agenda? As an outsider, I cannot know the relationship between the *intent* of the designer and the *encoding* of the design. But that relationship is key to high stakes accountability. If the authors of accountability reports are unaware that their design decisions impact audiences then there is a tremendous value to be gained in professional development. On the other hand, if authors do know the power of design, how successfully are they implementing their vision? Are authors making design choices that encourage their desired interpretations of the text? If not, there is again a strong argument for professional development. More design-literate authors will communicate their message more successfully. (Note that this question avoids asking whether the "desired" interpretation is good, bad, or neutral.) At the end of the day, these questions around the

designs and designers of accountability reports represent a gap in our current understanding of high stakes accountability.

Third, what are the impacts on audiences themselves? Are there real effects? Do different representations of the same information truly impact how audiences both interpret, and make use of, accountability data? If so, to what extent? Is the representation of some content more influential than the representation of others? Do these effects impact certain audiences more than others? More research is needed to untangle the complex relationship between the authors' intentions, the translation of intention into form, and the impact of form on consumers.

High-stakes accountability exists as a mechanism for the public to vet the quality of service provided by their tax dollars – to ensure that their children and their community are receiving the education they deserve, and if not, to exert direct influence on the schools themselves. In this model, accountability reports are the means of communication. They message to the public whether a school is meeting expectations, exceeding them, or falling short.

But, what if the message isn't clear? What if the message is delivered in such a way that some key points are emphasized more than others? Or what if the message is delivered in a way that frustrates, confuses, and demonstrably *misinforms* the audience? What if, despite best intentions, report authors make choices that muddy their messages? Suddenly the model falls apart. Rather than empowering the public to effect change, accountability reports might be complicit in reinforcing inequities and injustice within public education. As it stands, we simply do not know – a gap in understanding which may prove increasingly costly to public education nationwide – and one which demands additional investigation into the role that these reports play in public education.

CLARITY AND CONSISTENCY IN CONTEMPORARY STATE ACCOUNTABILITY REPORTS: EXAMINING THE DATA VISUALIZATIONS USED IN MANDATED PUBLIC SCHOOL REPORTING

Abstract

Since the passing of No Child Left Behind (NCLB) through to the current regulations of the Every Student Succeeds Act (ESSA), federal law mandates that all states receiving federal money to support public education must publish data about the performance of their schools. This article provides a survey of these reports. Drawing on data visualization research, I provide a content analysis of contemporary state-level accountability reports, comparing their design practices against the extant literature. The results show little consistency in either form or content of these reports. States largely favor designs that eschew graphical reporting in favor of tabular reporting, indicating a greater concern for precision and accuracy over context and comparison. Additionally, several states include design practices which, according to prior research, limit audiences' ability to accurately interpret and comprehend the information provided.

Introduction

Over the past five years, the political environment relating to public school accountability has been incredibly turbulent. From the original No Child Left Behind (NCLB) statutes, the Obama administration's NCLB-waiver system, and the passing of the Every Student Succeeds Act (ESSA), the legislative mandate requiring states to report on school quality has been in flux. Yet, despite this upheaval, one key regulation has remained: namely, that all 50 states are required to publish school accountability reports.

Within public school accountability legislation, these reports serve a critical role. Accountability reports are the primary mechanism for providing the public with information about the quality of public schools (Isaacs, 2003). Each report is, to some degree, a public-facing report card – a transparent document designed to show whether a given public school is meeting its obligation to provide a high-quality education to the students that it serves. This information is made available so that the public can hold schools accountable for their performance. Parents, families, and community members are expected to use the information contained in these reports to put pressure back on schools and to drive those schools towards improvement (Rogers, 2006).

As the regulatory landscape has changed, as the political definition of a "quality" education has shifted from administration to administration, states are forced to revise and republish their accountability reports. Often these changes are changes in content: for example, which measures to include, how much weight should be given to each, etc. The choice of what content to include in accountability reports is often dictated by the legislation itself; each new law includes slightly different provisions for what states must publish in their reports (McGuinn, 2016). However, the choice of what content to include is influenced by numerous factors, including not only the current law, but also the larger legacy of accountability legislation, as well as the individual motives of states themselves (Wong, 2015).

Along with the choice of content comes the choice of form. States must not only decide what information to report, but how to present that information to the public. As it turns out, unlike issues of content where states can often turn to the federal regulations themselves, there are few clear legislative directives when it comes to the display of information (ESSA, 2015; NCLB, 2001). States are encouraged to keep the information clear and concise, without specific guidance for how to do so ("State and

Local Report Cards: Title I, Part A of the Elementary and Secondary Education Act of 1965, as Amended, Non-Regulatory Guidance," 2013).

This lack of support is troubling, given the growing research in the field of data visualization. This research suggests that the choice of form is non-trivial. According to these researchers, the way in which authors represent information can impact how audiences will interpret that information (Cleveland, 1994; Tufte, 1997, 2006). The exact same content, presented to audiences in different formats, may lead audiences to very different understandings. Seemingly benign design choices – e.g., the use of tables vs. charts, the style and size of fonts, the inclusion of labels, axes, and legends, the use of white space – all have an impact on how audiences engage with the information that is displayed (Gribbons, 1992; Jacobsen et al., 2014; Shah et al., 1999). This research suggests that there is no such thing as a "neutral" design choice. Rather, every choice is a trade-off, prioritizing particular types of engagement with the text while obscuring others (Macdonald-Ross, 1977).

Given the importance of accountability reports to the broader system of school accountability, coupled with the lack of guidance given to states when designing these reports, there is a pressing need to support states on these complex issues of form. This research is a first step in that direction. In particular, this article attempts to evaluate states' current reporting practices, contextualizing actual report designs with existing research in data visualization. To do this, I begin by disentangling the complicated relationship between both form and content within contemporary accountability reports, looking not only at what content is included on these reports, but at how different states choose to represent the same content. Through a detailed content analysis, I examine what data visualization choices states are currently making, how these choices vary based on the content reported, as well as how these choices have varied over time within

individual states. With these data, I then apply key findings from data visualization literature, hoping to uncover what biases are implicit in current reporting practices and to point the way towards future research.

School Accountability Legislation

Broadly speaking, contemporary school accountability is inextricably tied to NCLB (2001). Prior to NCLB, federal oversight of public education in the United States focused primarily on inputs – on the distribution and quality of teachers, on the specifics of curricula enacted in the classroom. Through NCLB, this focus on inputs was replaced with a focus on outcomes – onto measurable markers of student learning (Isaacs, 2003; Wong, 2008). In addition to this mandated content, the legislation also required that all states publicly report this content via accountability reports. However, the NCLB statutes were far less prescriptive in form than in content. The legislation states that reports must be "concise [and] presented in an understandable and uniform format, and to the extent practicable, provided in a language that the parents can understand" (NCLB, 2001). There are no formal guidelines for the specific methods of presentation, or even best practices in terms of report design to which states must adhere.

As regulations changed through the Obama administration and into the Trump administration, the legacy of NCLB guidelines for both form and content remained roughly the same: legislation (or legislative guidelines) focused on content at the expense of format. For example, in response to states' criticisms of NCLB's incredibly high expectations (i.e., completely closing the achievement gap within 12 years), the NCLB-waiver system allowed states to set their own, more realistic goals, which would then be vetted by the Department of Education (McMurrer & Yoshioka, 2013). By definition, this approach was less strict than NCLB with regards to required content; however, the

Department of Education still provided several guidelines for what content to publish, including math and reading assessments and college- and career-readiness metrics. At the same time, the waiver system, like NCLB before it, provided few guidelines for reporting this information publicly. The only real guidance provided came in the format of a 2013 Department of Education guide to states. In this guide, the Department suggested that "an effective report card presents student and school performance data in a manner that is clear, easy to understand, and accessible to all stakeholders and, most especially, parents of the students who are the focus of ESEA programs" ("State and Local Report Cards: Title I, Part A of the Elementary and Secondary Education Act of 1965, as Amended, Non-Regulatory Guidance," 2013). There is no clear prescription on how to achieve these goals.

Today, requirements under ESSA echo those of the waiver era. States are required to submit personalized accountability plans, including indicators of college- and career-readiness that are aligned to high-quality summative assessments, with results broken down and reported by specific student groups (ESSA, 2015). At the same time, states are free to publish these data in whatever format they choose. As with the original NCLB legislation, ESSA requires accountability reports to be "concise, presented in an understandable and uniform format, and accessible to the public" (ESSA, 2015). And, as during the waiver era, the Department of Education provides loose guidelines for report designers, providing them with high-level questions to guide their design process (e.g., "Are the data available in both chart/graph and table format? Do the graphics and artwork improve readability and maintain user interest?") ("State and Local Report Cards: Title I, Part A of the Elementary and Secondary Education Act of 1965, as Amended, Non-Regulatory Guidance," 2013). However, the rest is up to the states themselves.

What this means in terms of contemporary accountability reporting is that states are permitted to vary widely in both the content presented and the form in which that content takes (Thomsen, 2013). Even during the NCLB era, when content was strictly mandated, states used the accountability reporting requirements as an opportunity to disseminate additional public information about their schools – whether that included other indicators of student performance, school inputs (e.g., staffing), or even school information (e.g., current principal, mission statements, etc.) (Wong, 2015). With the addition of the waiver program and the passing of ESSA, the room for variation in accountability design only increased. And, in fact, this variation is by design. The law is written to empower states to create their own individualized systems of accountability (McGuinn, 2016). This, however, is where the difference between content and form matters. The legislation gives states the flexibility in both the content and the form of accountability reports, but the content is regulated and the form is not. States can define their own measures of accountability if and only if those measures pass the Department of Education's formal review. At the same time, states can represent that content however they want, with no oversight whatsoever.

Parents' Engagement with School Reports

Although accountability reports have been a mainstay of public school accountability for over a decade, little research has focused on how parents engage directly with these documents. This is particularly troubling, given that parents' beliefs about their students' academic performance and parents' beliefs about school quality are often inaccurate. Recent work by the non-profit advocacy organization Learning Heroes (2017) demonstrates this disconnect quite profoundly. In a national survey of over 1,400 parents of public school children, Learning Heroes found that nearly 9 in 10 parents

believe their child is performing at or above grade level in math and reading, despite data showing only 1 in 3 students are actually doing so. At the same time, the majority of parents also believe that their students are receiving a high-quality education, with nearly 80% of respondents suggesting that the education their child is receiving is "Pretty good" or "Excellent".

Given this gap in expectation and reality, it is particularly important to understand how parents are engaging with information about school quality. The most recent and most direct research in this area comes from the Education Commission of the States' (ECS) review of accountability reporting in 2014. In this research, ECS asked a panel of parents to review sample report cards from all 50 states and to evaluate each report card on its level of accessibility and usefulness (Mikulecky & Christie, 2014). Overall, ECS found that parents heavily favor "report cards with clear graphics that made the data easy to understand" (p. 9). Although the ECS report does not provide summary information detailing what percent of states' reports met this criteria, the report does provide some guidelines for differentiating clear reports from opaque ones.

In particular, parents suggested that the easy-to-read reports: 1. included clear directions for interpreting the information, 2. provided clear presentations of information, and 3. avoided clutter and overwhelming detail. With regard to point one, clear directions, parents preferred reports that included instructions on how read the report itself (e.g., participants described liking reports that "provided directions as to how to navigate the page") as well as contextual information for the data displayed (e.g., participants disliked reports with "not much reference or explanation of the [data]"). With regard to clear information displays, parents preferred reports that included both tables and bar charts, as well as online reports that allowed easy access to additional data. With regard to clutter, parents consistently disliked reports that overwhelmed

readers with detail. Participants praised reports that "[were] not overwhelming with data" while criticizing reports that included academic jargon (e.g., "They use words that are not meaningful to the general public") and reports dense with detail (e.g., "Like reading a corporate financial report of 20 pages to get information").

Moving outside of accountability reporting, many researchers have examined how parents interact with other forms of academic data reporting. For example, early work by Aschbacher and Herman (1991) looks specifically at the design of student assessment reports, providing a comprehensive view of assessment reporting practices, as well as guidelines for creating effective assessment reports for parents and the public. In this work, the authors lay out several key recommendations for reaching parents: 1. know the audience and purpose; 2. keep it simple; 3. be clear, accurate, comprehensive, and balanced; and finally, 4. use techniques to capture and focus the reader's attention.

Elaborating on Aschbacher & Herman's call to "know the audience", Goodman & Hambleton (2004) remind report designers that public audiences often have little experience interpreting highly technical score reports. Echoing the ECS' research memo, they advocate for report designers to include explanatory text and/or supplementary information to walk readers through the appropriate interpretation of the report results. Similarly, Hambleton & Slater (1996) find that readers struggle with both the complex jargon and the technical statistics within assessment reports. Readers struggle with mathematical inequality symbols, references to "statistically significant" values and "standard errors", as well as cumulative score descriptions (e.g., "X% of students were at *or above* level Y"), Zwick et al. (2014) confirm that practitioners often struggle with statistical concepts and generally prefer "information presented as short, easy-to-read pieces." Reports with large blocks of dense text are often misinterpreted, if not completely ignored, by most readers. Finally, researchers emphasize the importance of

maintaining a consistent design. For example, Gribbons (1992) suggests that when section headers are clearly distinguished with consistent typography, when tables are formatted identically, and when charts share the same orientation, colors, and labels, readers will more easily navigate and interpret report information (1992). Similarly, Wainer (1997) argues that visual displays can maintain clarity by avoiding clutter, employing clear spacing, and ordering information in a reasoned way.

Across this work, clarity and consistency are key. Because the audiences of academic reporting are often inexperienced with interpreting complex score reports and educational jargon, designers must work to remove distractors and present information in a standardized way. Looking back to Aschbacher & Herman's (1991, p. 10) guidelines for effective score reporting, the message is plainly put: "Be consistent. Readers tend to resist changes in information representational styles."

Literature on Data Visualization

In addition to this research on parents' interpretation of academic data, there is a significant body of literature that examines how all audiences, regardless of context, make sense of data displays. According to this research, the way in which data reports are designed influences how audiences make sense of, and act upon, the data that are reported. Moreover, across this research, findings emphasize that there is no such thing as a "best" design. Rather, each design choice is a trade-off between competing goals and competing uses of the information displayed. One of the first, and most comprehensive, reviews of literature in this field comes from the work of Michael MacDonald-Ross' 1977 article "How Numbers Are Shown." In the article, MacDonald-Ross strongly emphasizes this notion of trade-offs:

> No one graphic format is universally superior to all others… To choose the best format for a particular occasion one must decide: what kind of data is

> to be shown? What teaching point needs to be made? What will the learner do with the data? (1977, p. 401)

Rather than thinking in terms of which design choice is better or worse, or what the "best practices" of design might be, MacDonald-Ross suggests pushing the analysis a step further by asking what choices might be better or worse for specific audiences when they are trying to achieve specific goals.

Tables Versus Charts

One of the most basic decisions facing report designers is what broad type of data visualization to use. Although the phrase "data visualization" often conjures images of brightly colored bar charts and infographics, the most straightforward (and perhaps most common) data visualization is a simple data table. Tables are particularly unique as a visualization because they are completely abstract – an ordered collection of labels and numbers, without any physical referents. Unlike a bar or line chart, which physically represents the differences in data points by length and by position, tables offer no such embodied representation of the data they contain (Macdonald-Ross, 1977). At the same time, this abstractness is paired by an extreme level of exactness. Every number is written in full, with as much precision as the author chooses to include (Tufte, 2001).

Comparatively speaking, graphical designs provide the exact opposite affordances of data tables. Charts and graphs are frequently less precise than data tables, but they more readily demonstrate differences in values (Schonlau & Peters, 2012). To a certain degree, this is intuitive. At best, a chart can be equally as precise as a data table, by including the exact legends, axes, and data labels one would find in a comparable data table. However, even without these supporting features, visual displays like bar and line charts allow readers to quickly and easily compare data points within a display (Jarvenpaa & Dickson, 1988). Participants make more accurate interpretations when the

difference between groups are expressed on a common scale or shown via length (Brinton, 1914; Cleveland & McGill, 1984; Karsten, 1923; Schonlau & Peters, 2012). Even without knowing the exact length of a bar or the exact position of a line on the y-axis, one can visually separate one bar from the next, one line from the other. This ability to visually differentiate the relative size of values is something which data tables inherently lack.

Consequently, the choice between table and chart involves compromise. "The choice of table," MacDonald-Ross writes, "involves a complex trade-off between compactness, exactness, and ease of usage" (1977, p. 379). Data tables are comparatively more precise and exact than charts and graphs, presenting more concrete pieces of data within the same space on a page (Tufte, 1997). Moreover, because data tables are so dense with information, they allow for exploration and investigation (Wainer, 1997). Through large, dense data tables, authors can include a multitude of information at once, creating a display which rarely prioritizes one single piece of information over any other, but instead allows readers to search for the information that is of most interest. Charts and graphs, on the other hand, sacrifice this exactness and compactness in favor of emphasizing differences and trends (Shah et al., 1999).

Micro Design Decisions

Research also emphasizes the importance of more granular design choices in maintaining accuracy and clarity. For example, Schriver (1997) and Vaiana and McGlynn (2002) argue for the importance of typographical styles and size in audiences' interpretations of data. Horton (1991) and Winn (1991) outline the importance of deliberate and judicious use of color to emphasize key points. Wainer (1997) emphasizes the need for white (or empty) space within displays. More broadly, Gribbons (1992)

describes how an array of choices including type, color, and spacing, as well as horizontal and vertical alignment, can cue the reader towards specific pieces of information. Moving to even more micro-design choices, researchers make a strong case for paying attention to fine-grained details of charts and graphs. “The basic elements,” writes Cleveland, “tick marks, scales, captions, plotting symbols, reference lines, keys, and labels… are critical controlling factors whose proper use can greatly increase the accuracy of the information that we visually decode from displays of data” (1994, p. 16). Tufte (1997) emphasizes the importance of tightly linking visual and textual elements; without clear labels and markers, charts become cumbersome to decode. Clearly depicted axes titles and chart labels help remove abstraction and improve legibility. In addition, Huff (1993) emphasizes the importance of graphical scale. When authors choose to truncate or otherwise manipulate the ordinate and abscissa (e.g., by plotting a chart from 80 to 100, rather than 0 to 100), the resulting image is misleading. By truncating axes, small changes in absolute values appear prominently; by stretching axes, large changes in absolute values are diminished.

Explanatory Text

Although the choice of data display and the micro-designs supporting that choice are critical, data visualizations do not exist in a vacuum. Instead, echoing parents’ feedback on school accountability reports, practitioners and researchers in the field of data visualization emphasize the fundamental need for explanatory text to support data designs themselves. This point is perhaps best voiced by data visualization expert Edward Tufte. In his book *The Visual Display of Quantitative Information,* Tufte writes, “Words and pictures belong together. Viewers need the help that words can provide [to understand the pictures they accompany].” (2001, p. 180). By combining each data

display with clear explanatory text, report designers can provide their audiences with context and framing for the information itself, leading to clearer understanding (Goodman & Hambleton, 2004; Yau, 2011). In the words of MacDonald-Ross, "[data visualizations] work best when accompanied by text. The text should not just repeat points made by the chart: it should direct, comment, explain, and question" (1977, p. 402). Moreover, these explanations are most valuable when they are clearly and directly tied to specific data visualizations. When explanatory text is printed immediately adjacent to a display (rather than on a separate page, or in a separate appendix, for example), readers can more easily connect this framing information to the display itself (Tufte, 2001, p. 181).

Inclusion, however, is the bare minimum. While research suggests that data visualizations can be improved by including explanatory text, the text itself matters. In order to be useful, explanatory text must be clear, meaningful, and accessible to report audiences. This point is well-established by research in the field of healthcare, specifically with regard to patient care and patient communication (Kelly & Haidet, 2007; Powers, 1988). In medicine, clear communication of health risks and treatment options is critical. Healthcare providers must ensure that they communicate in language that matches the literacy levels of their patients. As it turns out, often they do not (Cooley et al., 1995; Meade & Byrd, 1989). Similarly, within school accountability reports, it is imperative that the language of the reports matches the literacy level of audiences. Research suggests that the average adult in the US reads at approximately an 8th grade level (Bendick & Cantú, 1978; Doak, Doak, & Root, 1995; Eltorai, Sharma, Wang, & Daniels, 2015). Within the field of medicine, professional organizations like the National Institute of Health and the American Medical Association recommend that healthcare providers create written materials at a 3rd to 7th grade level (Hansberry, Agarwal,

Gonzales, & Baker, 2014). Within education, recommendations similarly vary between 5th and 9th grade levels (Nagro & Stein, 2016). Regardless of the exact cutoff, the point remains: explanatory text is only effective insofar as audiences can make sense of the text itself.

Applying This Research to Accountability Reports

Public school accountability is an enormous political machine that influences the lives, and future livelihood, of millions of American families. At the heart of this machine is the belief that communication will drive change, that if state departments of education provide administrators and parents with clear information about school quality, those administrators and parents will push for positive change (Shaul & Ganson, 2005; Wong, 2008). Yet, while the legislation provides states with guidelines on the types of data to include, states nonetheless have an incredible latitude in the choosing both *what* is presented and *how* it is presented. Furthermore, under ESSA, the *what* is explicitly vetted; the *how* is not. Once states decide what to report on, they must submit their proposals to the Department of Education for Review to ensure that they have chosen acceptable, high quality measures of accountability (Klein, 2016). No such review is required for the reports themselves.

This research is a first step in providing such a review. It is an attempt to catalogue the ways in which states are choosing to display school accountability data. The goal of this analysis is to understand the data visualization and design decisions currently used in accountability reporting and to see what types of audiences and readings these choices might inadvertently bias. In particular, by reviewing a sample of contemporary and historical state accountability reports, I hope to provide an answer to the following research questions:

1. What design elements are states using to report accountability data?
2. How have these design choices changed over time?
3. Based on existing research in the field of data visualization, what do these choices suggest about how accountability reports are shaping audiences' interpretation of school quality?

Method

To answer these questions, I conducted an in-depth content analysis of contemporary state-level public school accountability reports. Broadly speaking, this involved selecting several sample states for comparison. For each state selected, I collected a sample of the most recently published accountability reports accessible online. Additionally, reports from 5 and 10 years prior were also collected in some states to document changes in design over time. For all reports, I analyzed design choice using two separate approaches: first, by comparing how different states chose to represent the same content (e.g., examining how multiple states represent attendance and/or graduation rates), and second, by looking at broad design choices within each report, independent of content. Detailed descriptions of this method are provided below.

Sample Selection

Because the NCLB-waivers and the current ESSA legislation give states autonomy over both what to publish (i.e., content) and how to publish (i.e., form), I anticipated wide variation in both categories. To control for this variation in content, states were grouped based on the content that they publish. The source of information for this grouping was a publicly available, web-accessible database of accountability reports published by the Education Commission of the States (ECS) (Thomsen, 2013). The ECS

database, published in 2013, provides the most up-to-date catalog of state accountability reporting, including a detailed list of content included in each state's report. The majority of reports in the database were published for the 2011-12 school year. As a result, the ECS database reports are more likely to reflect original NCLB requirements than those of NCLB-waiver flexibility. Within the database, each of the 50 states are represented, along with the District of Columbia, American Samoa, Guam, Puerto Rico, and the Virgin Islands. For each, the database provides a standardized list of measures that are included on the report. A summarized list of these measures is included below.

Table 1. ECS metrics reported by state (Thomsen, 2013)

Metric Name	Type	Total Reporting	Percent Reporting
Attendance Rate – Secondary	General	23	41.82
Teachers - % Highly Qualified	General	23	41.82
Enrollment	Profile	24	43.64
Annual Measurable Objective AMO or AYP	General	29	52.73
Attendance Rate – Elementary / Middle	General	31	56.36
Student Demographic / Socioeconomic Data	Profile	33	60.00
Growth / Academic Progress	General	36	65.45
Achievement Gap Closure	General	40	72.73
Graduation Rate	General	53	96.36
Assessment Scores / Student Achievement	General	55	100.00

Using this database, states were split into "reporting groups" such that each group consisted of states reporting similar types of content to one another, but significantly different content than states in other groups. Specifically, an exploratory factor analyses was conducted to test for the presence of hidden or "latent" variables which might explain differences in metrics reported across states (Fabrigar, Wegener, MacCallum, & Strahan, 1999). During the factor analysis, a handful of factors emerged which held up to both statistical and analytical scrutiny. Ten factors were found with eigenvalues greater than two; traditionally, the larger the eigenvalue, the more observable variables are explained by a single latent variable (a rule of thumb is to include factors with eigenvalues greater than 1). Additionally, when examined analytically, each of these factors maintained face validity. For example, factor two loads heavily on the presence/absence of GED passage rates and WorkKeys reporting (an ACT-administered job skills assessment), pointing to a latent variable tied to student employment reporting. Similarly, factor three loads heavily on the reporting of class size, facilities expenditures per pupil, and institutional and curricular materials, pointing to a latent variable tied to reporting on classroom inputs.

Using these factors, reporting groups were created via a cluster analysis using Ward's method. Based on the fit statistics, a 9-cluster solution was chosen, providing three groupings of states (n=19, 9, and 17, respectively) and six smaller clusters (n=10). Based on these analyses, there was good reason to think that states do, in fact, fall into discrete reporting groups. The 19 states in Cluster 1 are distinguished by the absence of reporting on a factor associated with attendance rates in elementary, middle, and secondary schools. Although there is variation in the content reported by these states, they are more likely than other states to *exclude* content related to student attendance. By contrast, the 17 states in Cluster 3 are distinguished by the presence of this factor (i.e.,

states that are more likely to report on attendance); however, these states are also associated with the absence of reporting on the factor tied to ACT/SAT participation rates and AP participation rates and scores. Finally, the 9 states in Cluster 2 are marked by the presence of a factor tied to ACT, SAT, and AP reporting, as well as a factor tied to college attendance rates and IB participation. In other words, these schools represent a cluster that is more likely to report on college-readiness programs and college attendance outcomes.

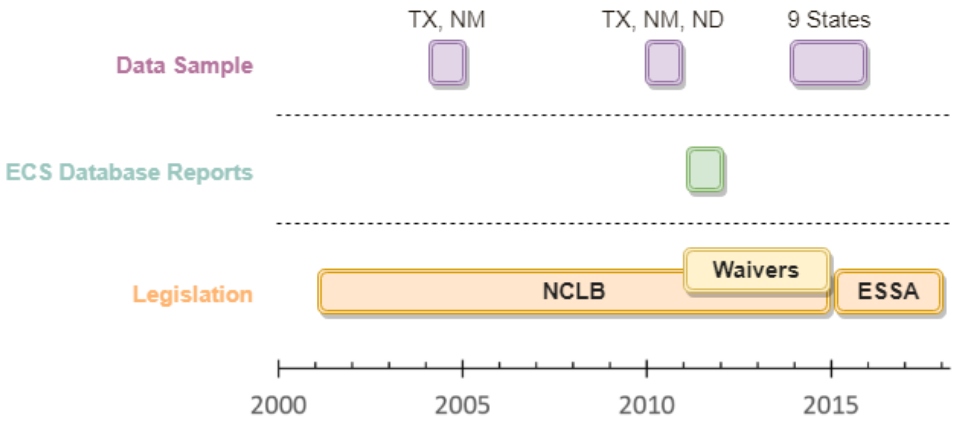
Table 2. Cluster analysis and factor analysis summary

Cluster	States	Associated Factors	Metrics Tied to Factor
Cluster 1	19	Absence of Factor 9	Attendance Rates
Cluster 2	9	Presence of Factor 6	ACT/SAT participation; AP participation and scores
		Presence of Factor 8	College going rate; IB participation
Cluster 3	17	Absence of Factor 6	ACT/SAT participation; AP participation and scores
		Presence of Factor 9	Attendance Rates
All Others	10	-	-

From each of these clusters, 3 representative states were selected at random: from Cluster 1, Texas, Maryland, and Maine; from Cluster 2, New Jersey, New Mexico, and Oregon; from Cluster 3, North Dakota, North Carolina, and Wisconsin. Within this sample of 9 states, the most recently available high-school level PDF accountability reports were found for each state using each state department of education's publicly accessible website. The sample was limited to high school reports based on an initial survey of reports which showed little variation in reporting between high school and elementary grades, as well as a propensity for high school reports to include more

information about students and schools (e.g., graduation rates, college & career measures). Data were collected for the 2014-15 school year for all states except Maine and Oregon, where the 2015-16 reports were available (herein referred to collectively as the 2014-16 reports). Additionally, one state from each cluster was selected at random for a longitudinal analysis (Texas, New Mexico, North Dakota, respectively). For these states, accountability reports were also collected from approximately 5 and 10 years prior, where available. The historical data collected included reports for Texas (2004-05 and 2010-11 school years), New Mexico (2004-05 and 2010-11 school years), and North Dakota (2010-11 school year). No earlier reports were found for North Dakota. Finally, after all first-round analysis and coding was complete, an additional sample of three states' 2014-16 reports were collected to serve as a validity test for initial findings. These states, also chosen at random, included Missouri, Mississippi, and South Dakota. All reports, contemporary and historical, were accessed online in January of 2017, with the exception of the validity sample, which was collected in February of 2018.

Figure 1. Timeline of accountability legislation, ECS database reporting, and data sample.



Data Collection & Analysis

Within these sample reports, data were collected and analyzed using a syntactical approach (Krippendorff, 2012). In this model, documents are broken down into small, discrete units of syntax: e.g., text, charts, tables. Similarly, within each of these syntactical units, representations are further broken down. For example, charts have their own unique syntactic elements, including titles, axes, axis labels, data points, data labels, etc. This syntactical approach provides a reliable method of identifying data by providing easily distinguishing elements of content (Krippendorff, 2012). Additionally, these units of syntax align well to the elements described by existing literature. Data visualization research focuses concretely on the differences between charts and tables, on the inclusion of specific chart elements like axes, labels, and tick marks. By narrowing one's focus to these units of syntax, one can more easily apply existing theory and existing analytical frames to the collected data. Using this syntactical method, three different approaches to data collection and analyses were applied.

Pre-identified content. For a first stage of data analysis, the ECS database of accountability reports was used to identify common information reported across multiple states (i.e., measures listed above in Table 1). In addition to these commonly included measures, three others were selected (SAT/ACT, AP/IB, and College Going) based on the cluster analysis above, which suggested that states can be differentiated based on whether they include these measures in their reports. Using the sample PDFs, a first round of analysis documented whether these measures were included in each state's report. Next, for the measures that were reported, a second round of analysis identified the units of syntax used to display those measures: i.e., text, table, chart, or combination thereof. Depending on what unit of syntax was used, additional and more detailed information was captured on the more granular units of syntax, including, for example,

the choice of chart used (e.g., bar, pie, line), the inclusion of elements in the chart (e.g., legends, data labels), the styling of data tables (e.g., gridlines, font sizes). In addition, information was captured regarding the context in which data were displayed, looking specifically at the inclusion/exclusion of explanatory text adjacent to the data themselves. Finally, in instances where this text was included, analysis was conducted to evaluate the estimated reading level required to decode the text, using the Flesch-Kincaid (1975) and Coleman-Liau (1975) algorithms.

Holistic report review. In the second stage of data analysis, I captured variation in form independent of the content report. In contrast to the approach above, this holistic approach to data collection examined units of syntax at the report level, documenting the rates at which reports included or excluded various design elements. First, to see how frequently charts and graphs were used, I calculated the proportion of pages within each report that included at least one or more chart. This calculation was chosen to control for the variation in the amount of information published by each state (e.g., reports ranged from 2 pages to over 50 pages in length). Additionally, data was collected on the consistency of visual elements across units of syntax. For each report, binary measures were created to identify whether these units of syntax were present or absent for all charts of a specific type (e.g., across all stacked bar charts in the report, were min/max labels included on axes?). For reports that included multiple examples of the same units of syntax (e.g., multiple charts, multiple data tables), data were collected to capture consistency (or lack thereof) in color, size, alignment, and use of white space.

Results

Content

In the contemporary 2014-16 reports, there was significant variation in the content reported across states. The median page length of current year reports was 6 pages, with the shortest report including 3 pages of information and the longest report including 54 pages of information. As seen in Table 3 information about student assessment scores and student graduation rates were reported in all nine states, with demographic and attendance data following close behind (seven and six states reporting, respectively). Fewer than half of the states reported on any of the remaining content areas. Of the nine states, Maryland and Maine reported on the fewest of the observed content areas (three each), while New Jersey and New Mexico reported on the most (seven each). Every state reported on additional content beyond these areas of interest. However, no state's 2014-16 reports included measures of Average Yearly Progress (AYP) or Annual Measurable Objectives (AMO).

On the whole, sample states were more likely to include comparison information than to omit it. Across all states, nearly three out of four reported metrics included a comparison to either a state or a district average; however, sample states reported state averages nearly twice as frequently as district averages (81.4% vs. 38.9%). Inclusion of comparisons by student subgroup were also more common than not, occurring in 63% of cases, while comparison to prior years were included alongside just under half of the observed metrics.

Table 3.

Metrics reported by state in contemporary reports (2014-16)

	Cluster 1			*Cluster 2*			*Cluster 3*		
Metric	MD	ME	TX	NJ	NM	OR	NC	ND	WI
Achievement Gap Closure				x	x			x	x
AMO or AYP									
AP / IB				x	x				
Assessment Scores / Student Achievement	x	x	x	x	x	x	x	x	x
Attendance Rate	x		x		x		x	x	x
College Going				x					
Graduation Rate	x	x	x	x	x	x	x	x	x
Growth / Academic Progress					x	x	x		x
SAT / ACT			x	x	x		x		
Student Demographic / Socioeconomic Data		x	x	x		x	x	x	x

Among the historical reporting sample, for the 2010-11 school year, all three states reported on assessment results, graduation rates, and attendance measures. No state reported on college and career going measures or on achievement gap measures. Moreover, only one state, North Dakota, reported on an AYP/AMO measure. In the 2004-05 school year, both Texas and New Mexico included assessment results, graduation rates, and attendance measures, but did not include college and career going measures or achievement gap measures. Comparison data was also often included in prior reports. In 2010-11, just under three quarters of all measures included comparison data, while in 2004-05 all states' reported measures included comparison data. In both

years, comparisons to state averages were more likely than comparisons to district averages.

Overall, these findings vary significantly from the predictions of the ECS database and the subsequent factor analysis. The content included in sample states' 2014-16 reports does not neatly align with the content reported in the reports of the ECS database in which, for example, Texas would have been expected to not include attendance measures, and New Mexico would have been expected to report college- and career-going measures. There is little consistency between the content reported within states over time or across states within the same pre-identified reporting groups. That said, because the current analysis attempts to control for variation in content while analyzing choices of form and design (i.e., examining how states reported the exact same pieces of content), there is still solid ground for understanding reporting practices across states. In other words, although one cannot compare the differences between design choices in Cluster 1, Cluster 2, and Cluster 3, one can evaluate differences in how states report on common measures including assessment results, graduation rates, and demographic information, as well as on their overall tendencies in reporting independent of measure (e.g. percent of pages with or without charts).

Tables Versus Charts

Pre-identified content. Within the pre-identified areas of content, there was a clear tendency for states to report measures in data tables rather than in charts and graphs. As seen in Table 4, for the pre-identified content that states did report, eight of nine states reported every single measure as a table. The one exception, Oregon, reported all but one of its measures as a table. At the same time, there was an equally strong tendency for states to avoid reporting content in charts and graphs. Less than 14% of the

data visualizations featuring pre-identified content measures were reported as a chart or graph. Five of the nine states did not report any of the sample content with charts. Of the remaining four states, both New Jersey and Wisconsin reported a single measure with a chart, while New Mexico and Oregon deviated from the trend, reporting 43% and 50% of the pre-identified measures as charts. These findings also held true for the historical reporting sample. In 2010-11, all instances of pre-identified content were reported as a table; none were displayed using a chart or graph. In 2004-05, the same trend held.

Table 4.

Tables and Charts by Metric and State in Contemporary Reports (2014-16)

State	Metrics Reported	Number Reported as Table	Percent Reported as Table	Number Reported as Chart	Percent Reported as Chart
MD	3	3	100%	0	0%
ME	3	3	100%	0	0%
NC	6	6	100%	0	0%
ND	5	5	100%	0	0%
NJ	7	7	100%	1	14%
NM	7	7	100%	3	43%
OR	4	3	75%	2	50%
TX	5	5	100%	0	0%
WI	6	6	100%	1	17%

Overall, these results remained consistent on a metric-by-metric basis, though, there were some exceptions. While all states included state assessment data content and all states represented that content in data tables, one third of states also represented that same content in chart form (specifically, as bar charts). As a result, assessment data was

the content most frequently visualized as a chart. Demographics were shown in chart format in two states (Oregon, New Jersey), while AP and SAT were charted in just one state each (New Jersey). None of the other metrics (e.g., graduation rate, attendance rate, college & career going) were shown as a chart. Of all charts included (n=12) in the pre-identified content for the nine 2014-16 reports, bar charts were displayed most frequently (n=5), followed by stacked bars (n=3). New Mexico was the only state to include a combination of data table and heat map. New Jersey was the only state to represent a single piece of content in more than two forms, displaying their demographics as a table, bar chart, stacked bar chart, pie chart, and line chart.

Holistic report review. Looking beyond the pre-identified content areas, and considering each report as a whole, a similar preference for tables over charts emerged. Two states, Maine and North Dakota had a chart-inclusion rate of 0%: no charts were included on any pages of these states' reports for any areas of content. In contrast, two other states, Maryland and Oregon, had a rate of 50%, showing at least one chart or graph on every other page of the report. All other states fell between these two extremes, with a median rate of 33.33%. Across all states' 2014-16 reports, nearly 1 in 4 pages included a chart or graph, up from 1 in 5 pages for the 2010-11 sample. Of the charts that were included, bar charts were most frequently used, appearing on 13.6% and 16.5% of pages in current reports and 2010-11 reports, respectively. Stacked bar charts were the next most frequently used, appearing in nearly 9% of pages, while line charts and pie charts were less frequently used, appearing in fewer than 4% of pages. Looking historically, we see that this modest inclusion of charts and graphs is an improvement: in the 2004-05 sample, no charts or graphs were included on any page.

Examples. Because all states reported on end-of-year assessment results, and because states presented this information as both tables and charts, assessment results

provide a common ground for contrasting states' visualization choices. Although all states reported end-of-year assessment data in tabular format, states varied in the amount of information displayed as well as the presentation of that information. For example, North Dakota's report illustrates the traditional tabular display of assessment results, with columns reporting the percentage of students achieving each of the four performance levels on the assessment (i.e., novice, partially proficient, proficient, and advanced), while the rows of the table provide breakouts for each student subgroup (e.g., male, female, etc.) as well as comparisons to the district and state averages (see Figure 2). Similarly, North Carolina's reports include a tabular display of assessment data, with disaggregation by subject (i.e., English II, Math I, Biology) and performance level, also accompanied by district and state averages. However, unlike the North Dakota report, North Carolina displays the results for each performance level as a separate table, with its own accompanying explanatory text. Hearkening back to the ESSA reporting guidelines, the North Dakota report is more concise, with space for over 231 unique numbers on this single page, as compared to the 45 numbers displayed on North Carolina's report. However, the North Carolina report provides much greater narrative summary to support the report's readability and users' interest in the material. Descriptions are included for each of the five performance levels, both above each table and at the right of the report, along with additional text describing why these levels matter and how these levels were calculated. No descriptions are included on the North Dakota excerpt (all summary information is limited to a single explanatory page at the beginning of the report).

Figure 2. Assessment results from North Dakota's 2015-16 report

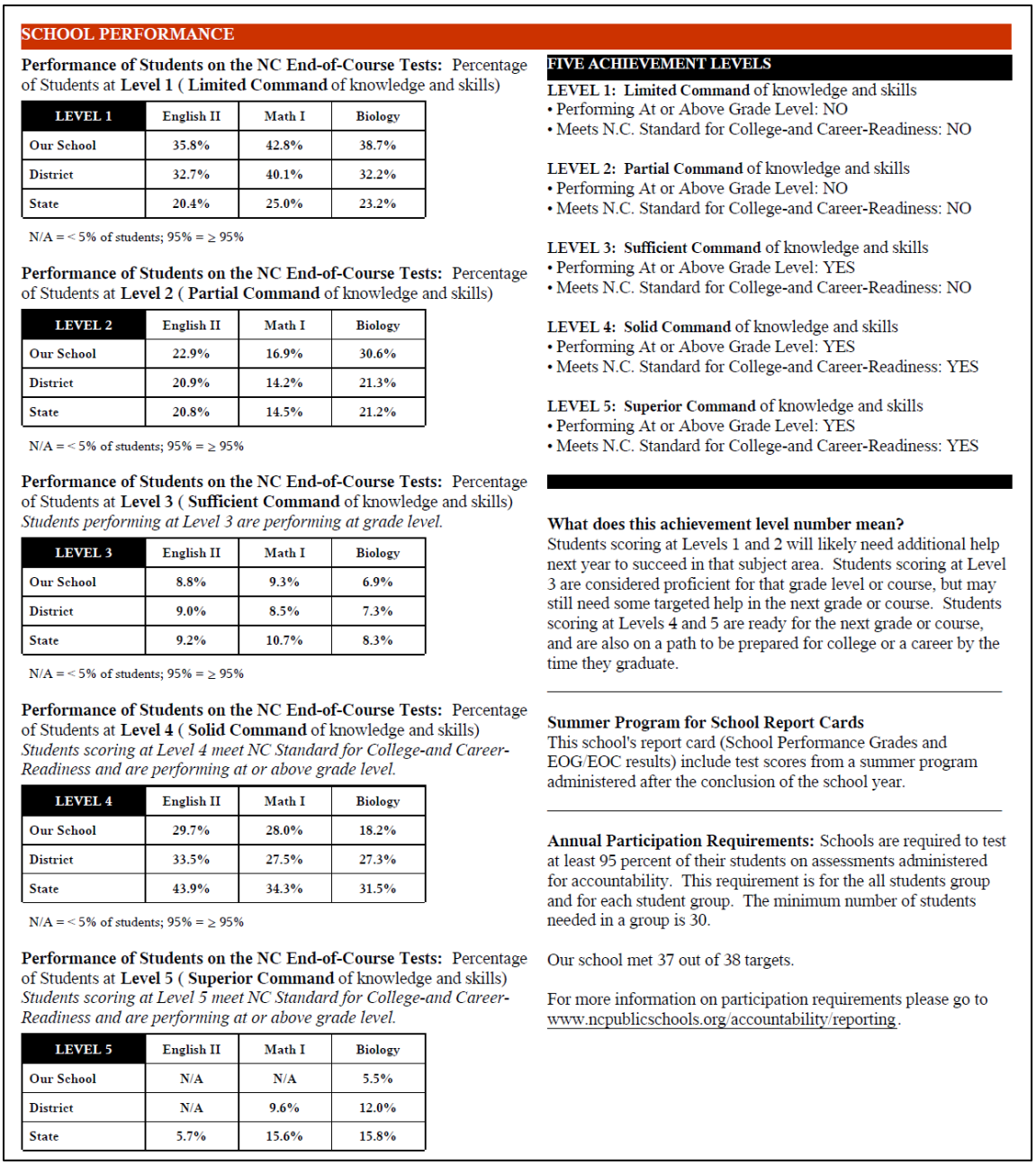
Report: North Dakota Assessment - School, District, and State **2015-16**

School: Starkweather Public School *36-044-8230-0712* **Starkweather 44 (PK-12)**

Math Achievement Rates (Across All Grades)[1] **Section C**

Group	Total Number of Students	Achievement Levels				Combined Levels	
		Novice	Partially Proficient	Proficient	Advanced	Not Proficient[2]	Proficient[3]
School - All	11	54.5%	18.2%	27.3%	0.0%	72.7%	27.3%
District - All	11	54.5%	18.2%	27.3%	0.0%	72.7%	27.3%
State - All	22000	29.2%	33.0%	24.8%	13.0%	62.2%	37.8%
School - Male	3						
District - Male	3						
State - Male	11211	31.7%	31.4%	23.3%	13.6%	63.1%	36.9%
School - Female	8						
District - Female	8						
State - Female	10789	26.6%	34.7%	26.2%	12.4%	61.4%	38.6%
School - White	11						
District - White	11						
State - White	17720	23.7%	33.8%	27.8%	14.7%	57.5%	42.5%
School - Native American	0						
District - Native American	0						
State - Native American	1996	58.1%	29.1%	9.7%	3.2%	87.2%	12.8%
School - Black	0						
District - Black	0						
State - Black	989	54.3%	30.1%	10.1%	5.5%	84.4%	15.6%
School - Hispanic	0						
District - Hispanic	0						
State - Hispanic	814	47.2%	33.3%	14.4%	5.2%	80.5%	19.5%
School - Asian American	0						
District - Asian American	0						
State - Asian American	481	31.2%	27.4%	21.0%	20.4%	58.6%	41.4%
School - Limited English Proficient (LEP)	0						
District - Limited English Proficient (LEP)	0						
State - Limited English Proficient (LEP)	547	>79.8%	15.2%	<4.1%	0.9%	>95.0%	<5.0%
School - Non LEP	11						
District - Non LEP	11						
State - Non LEP	21453	27.9%	33.5%	25.3%	13.3%	61.4%	38.6%
School - Low Income	5						
District - Low Income	5						
State - Low Income	6497	47.2%	32.5%	15.3%	5.1%	79.7%	20.3%

Figure 3. Assessment results from North Carolina's 2015-16 report



SCHOOL PERFORMANCE

Performance of Students on the NC End-of-Course Tests: Percentage of Students at **Level 1** (**Limited Command** of knowledge and skills)

LEVEL 1	English II	Math I	Biology
Our School	35.8%	42.8%	38.7%
District	32.7%	40.1%	32.2%
State	20.4%	25.0%	23.2%

N/A = < 5% of students; 95% = ≥ 95%

Performance of Students on the NC End-of-Course Tests: Percentage of Students at **Level 2** (**Partial Command** of knowledge and skills)

LEVEL 2	English II	Math I	Biology
Our School	22.9%	16.9%	30.6%
District	20.9%	14.2%	21.3%
State	20.8%	14.5%	21.2%

N/A = < 5% of students; 95% = ≥ 95%

Performance of Students on the NC End-of-Course Tests: Percentage of Students at **Level 3** (**Sufficient Command** of knowledge and skills)
Students performing at Level 3 are performing at grade level.

LEVEL 3	English II	Math I	Biology
Our School	8.8%	9.3%	6.9%
District	9.0%	8.5%	7.3%
State	9.2%	10.7%	8.3%

N/A = < 5% of students; 95% = ≥ 95%

Performance of Students on the NC End-of-Course Tests: Percentage of Students at **Level 4** (**Solid Command** of knowledge and skills)
Students scoring at Level 4 meet NC Standard for College-and Career-Readiness and are performing at or above grade level.

LEVEL 4	English II	Math I	Biology
Our School	29.7%	28.0%	18.2%
District	33.5%	27.5%	27.3%
State	43.9%	34.3%	31.5%

N/A = < 5% of students; 95% = ≥ 95%

Performance of Students on the NC End-of-Course Tests: Percentage of Students at **Level 5** (**Superior Command** of knowledge and skills)
Students scoring at Level 5 meet NC Standard for College-and Career-Readiness and are performing at or above grade level.

LEVEL 5	English II	Math I	Biology
Our School	N/A	N/A	5.5%
District	N/A	9.6%	12.0%
State	5.7%	15.6%	15.8%

FIVE ACHIEVEMENT LEVELS

LEVEL 1: Limited Command of knowledge and skills
• Performing At or Above Grade Level: NO
• Meets N.C. Standard for College-and Career-Readiness: NO

LEVEL 2: Partial Command of knowledge and skills
• Performing At or Above Grade Level: NO
• Meets N.C. Standard for College-and Career-Readiness: NO

LEVEL 3: Sufficient Command of knowledge and skills
• Performing At or Above Grade Level: YES
• Meets N.C. Standard for College-and Career-Readiness: NO

LEVEL 4: Solid Command of knowledge and skills
• Performing At or Above Grade Level: YES
• Meets N.C. Standard for College-and Career-Readiness: YES

LEVEL 5: Superior Command of knowledge and skills
• Performing At or Above Grade Level: YES
• Meets N.C. Standard for College-and Career-Readiness: YES

What does this achievement level number mean?
Students scoring at Levels 1 and 2 will likely need additional help next year to succeed in that subject area. Students scoring at Level 3 are considered proficient for that grade level or course, but may still need some targeted help in the next grade or course. Students scoring at Levels 4 and 5 are ready for the next grade or course, and are also on a path to be prepared for college or a career by the time they graduate.

Summer Program for School Report Cards
This school's report card (School Performance Grades and EOG/EOC results) include test scores from a summer program administered after the conclusion of the school year.

Annual Participation Requirements: Schools are required to test at least 95 percent of their students on assessments administered for accountability. This requirement is for the all students group and for each student group. The minimum number of students needed in a group is 30.

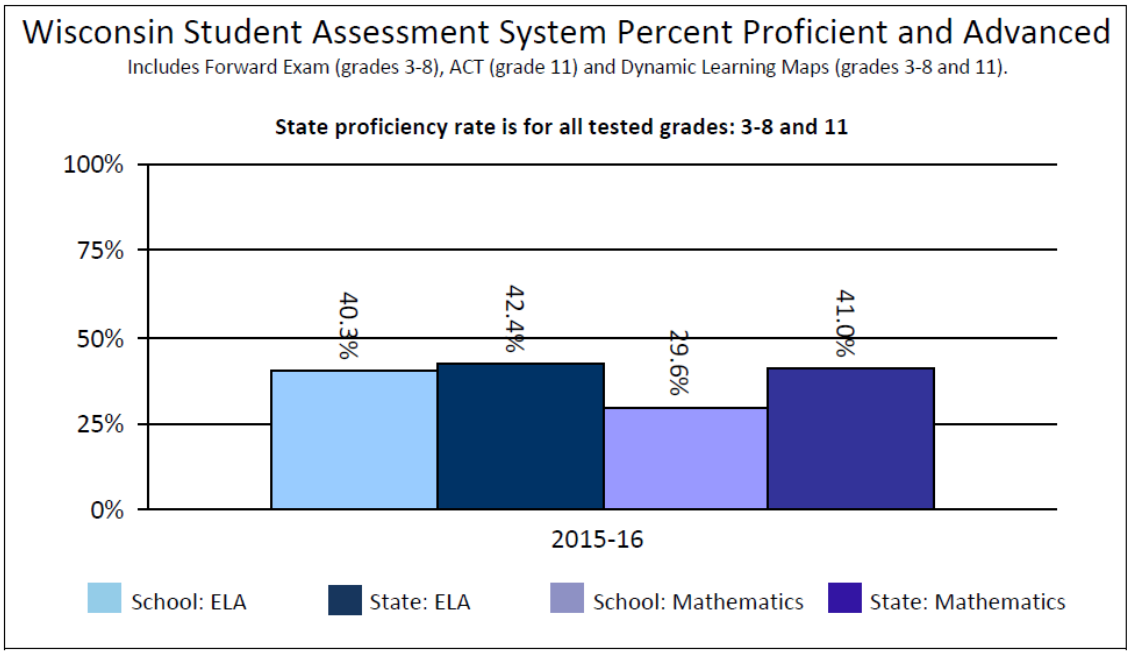
Our school met 37 out of 38 targets.

For more information on participation requirements please go to www.ncpublicschools.org/accountability/reporting.

In contrast to these traditional tabular displays, the Wisconsin accountability report includes a graphical representation of end-of-year assessment results. Figure 4 shows student performance on the state assessment in both Mathematics and English Language Arts (ELA), with additional comparisons to the state average. These results are

plotted on a traditional bar chart, with vertical bars sized to match the corresponding proficiency rates. Exact figures for each proficiency rate are included on the chart as data labels and each bar is additionally identified by color, with a legend included directly beneath the chart.

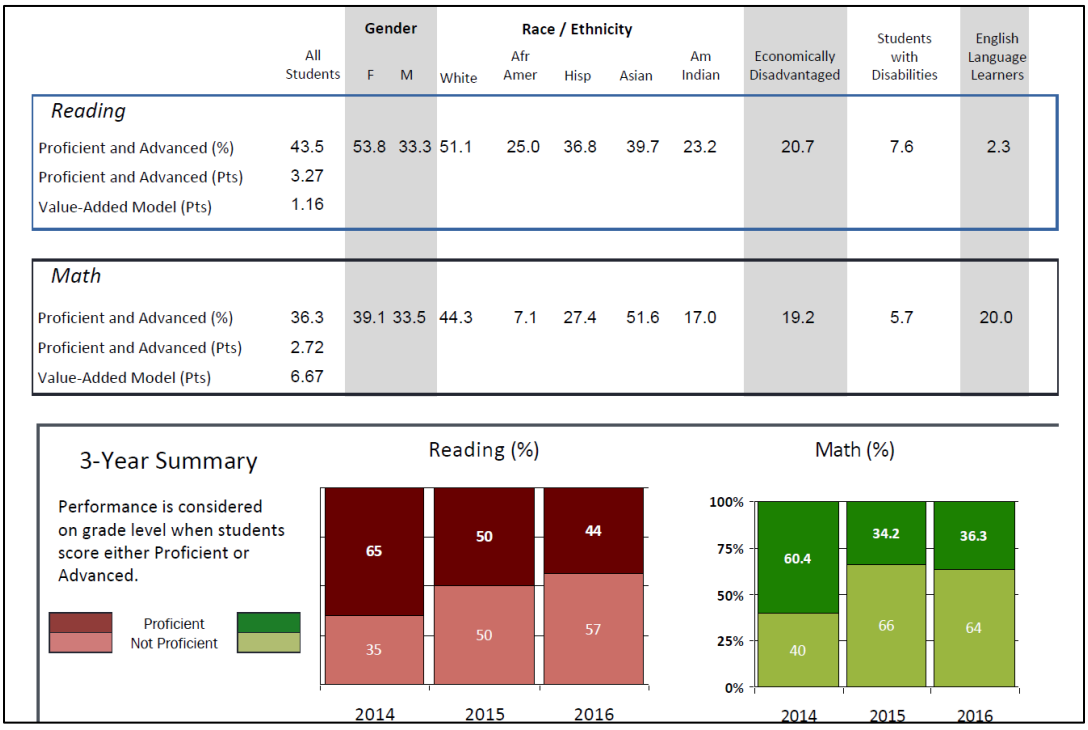
Figure 4. Assessment results from Wisconsin's 2015-16 report



Generally speaking, the use of charts (in addition data tables) improved over time. For example, New Mexico's 2015-16 report displays 3 years of accountability data in a stacked bar chart (see Figure 5). This display, featured directly beneath a traditional data table, shows the proportion of students who were proficient and who were not proficient, by subject, in 2014, 2015, and 2016, respectively. As with Wisconsin, data labels are included on the chart itself and bars are color-coded with a legend directly adjacent. By way of contrast, the 2010-11 New Mexico reports did not include any graphical representation of assessment data (see Figure 6). Instead, all data were

reported in a traditional data table with rows and columns closely matching those of the 2015-16 report (with distinctly different visual styling, however).

Figure 5. Assessment results from New Mexico's 2015-16 report



	All Students	Gender F	Gender M	Race / Ethnicity White	Afr Amer	Hisp	Asian	Am Indian	Economically Disadvantaged	Students with Disabilities	English Language Learners
Reading											
Proficient and Advanced (%)	43.5	53.8	33.3	51.1	25.0	36.8	39.7	23.2	20.7	7.6	2.3
Proficient and Advanced (Pts)	3.27										
Value-Added Model (Pts)	1.16										
Math											
Proficient and Advanced (%)	36.3	39.1	33.5	44.3	7.1	27.4	51.6	17.0	19.2	5.7	20.0
Proficient and Advanced (Pts)	2.72										
Value-Added Model (Pts)	6.67										

Figure 6. Assessment results from New Mexico's 2010-11 report

Alamogordo High

- Too few students to be reported	All Students	Gender F	Gender M	Race / Ethnicity White	Afr Amer	Hisp	Asian	Am Indian	Economically Disadvantaged	Students with Disabilities	English Language Learners
Enrollment (%)	100	49	51	50	8	36	3	3	47	13	1
Participation Rate Assessments (%)	100	100	100	100	100	100	100	100	100	100	-
Reading											
Status (% Proficient)	60.7	70.9	50.0	69.4	51.5	51.4	54.5	45.5	50.8	19.6	-
Growth Highest 75%	-	-	-	-	-	-	-	-	-	-	-
Growth Lowest 25%	-	-	-	-	-	-	-	-	-	-	-
School Growth	(Available in 2012)										
Math											
Status	42.4	43.3	41.3	51.0	36.4	30.4	54.5	36.4	32.4	13.0	-
Growth Highest 75%	-	(N/A)	(N/A)	-	-	-	-	-	-	-	-
Growth Lowest 25%	-	(N/A)	(N/A)	-	-	-	-	-	-	-	-
School Growth	(Available in 2012)										

Micro-Design Decision

Pre-identified content. There was little intra-report consistency among data displays for pre-identified areas of content. Measures of consistency were collected on the use of font, text alignment, white space, and color across each of the nine 2014-16 reports. On average, the use of font was most likely to be consistent from one element to the next, with two-thirds of states having consistent font faces and font sizes. Font sizes varied across the sample states, with a minimum of 6.5 points and a maximum of 12 points. On average, demographic measures were smallest on the page, with an average font size of 8 pts, while measures of academic growth and college and career readiness were largest on the page, with an average of 10 pts or larger. Beyond the fonts themselves, there was also significant variation in states use of white space, text alignment, and color. Only a third of sample states demonstrated consistent white space and consistent use of text alignment, and only two of the sample states were consistent in their use of color (e.g., in table headers and cell highlights and in bar and line shading).

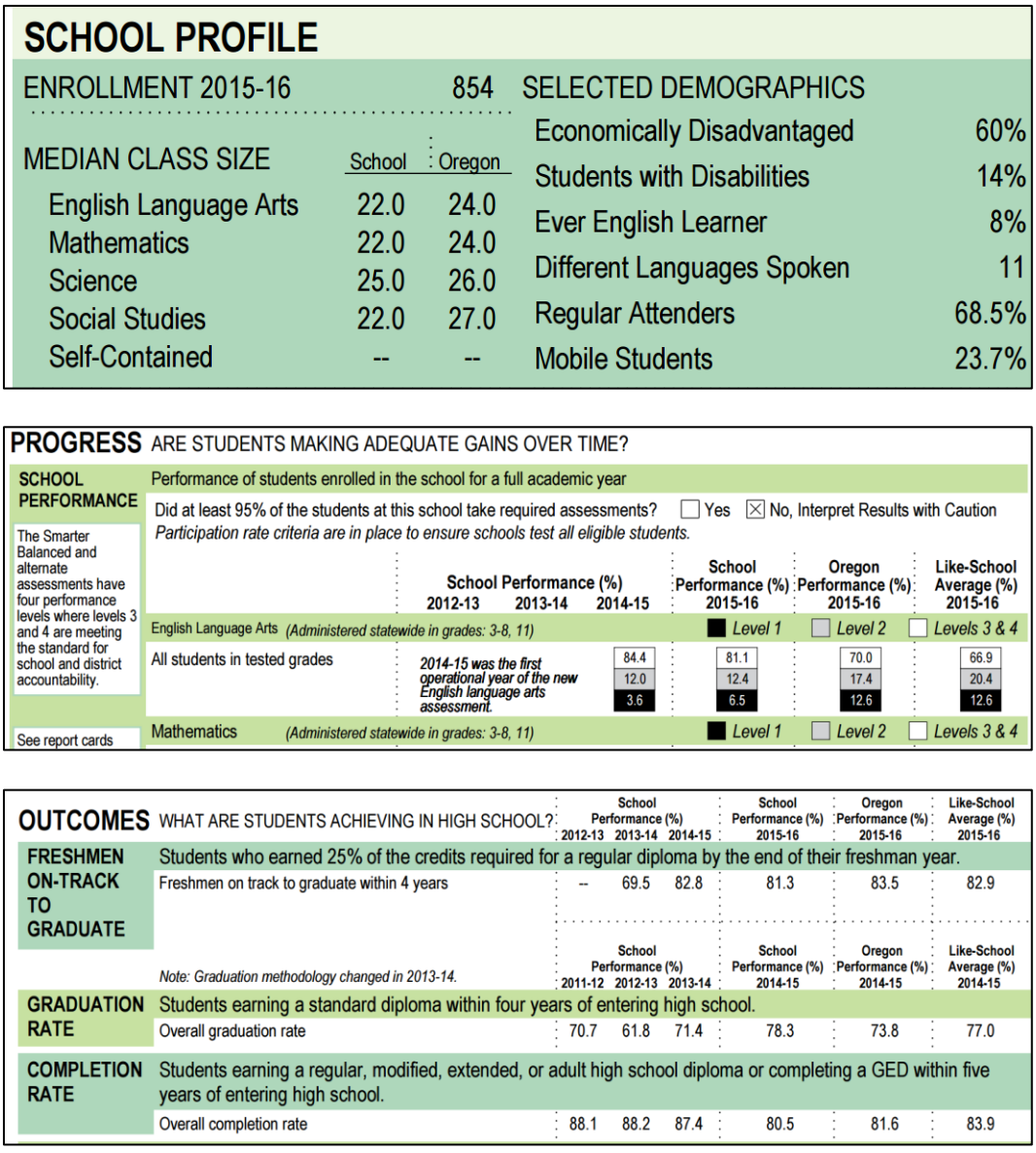
Holistic report review. Looking across all content areas in the report, there was a slight improvement in clarity and consistency between historical reports and contemporary reports. Based on the literature, a key aspect of clarity for charts and graphs involves consistent use of more granular units of syntax (e.g., inclusion of minimum and maximum labels on axes, clear data labels, and exact figures for each data point). Across the nine sample states, 2014-16 year reports included more of these clear and consistent syntactical elements than prior year reports. Looking by state and by chart type, legends were included over 90% of the time in contemporary reports, up from only 60% in 2010-11. Clearly labeled minimum and maximum axes values were included 80% of the time versus 20% in 2010-11. Data labels were included 76% of the time, and those labels were placed directly beside their data markers 70% of the time, up from 50%

for both in 2010-11. Finally, current year charts were more likely to be precise and accurate than prior reports. Across all 2010-11 reports (i.e., reports from TX, NM, ND), visible errors were present in 1 out of 4 chart types by state, and only half included precise values for the data displayed. In 2014-16 reports (i.e., reports from all nine sample states), error rates were cut in half, appearing in 1 out of 8 instances, while precise values were displayed in over 90% of cases.

Examples. The 2015-16 report card for Oregon is one of the shortest reports, with only 4 pages of content, yet across these four pages, not a single design element is repeated. Every section is distinct, and even within sections, there are often competing visualizations. Figure 7 includes four excerpts from the report, each using the same broad unit of syntax (i.e., a data table) yet with very different implementations. Across all four, the color scheme and fonts remain relatively consistent (though the first table is completely shaded, while others are only partially so); however, the layout is completely different. Across the section headers, notice that the first excerpt (i.e., "SCHOOL PROFILE") does not have a description, while the others do. Similarly, each header is separated from the table itself, with the exception of the third table, where the header presses against the column names. The left side of the table frequently contains a sub-section header, with a description written out along the full width of the table, except in the first table, where subsections are placed within the table itself, and in the second table, where additional descriptions are included in a white text box (a choice which is not repeated anywhere else in the report). The column names in the second and third table are identical; however, the font sizes and column widths vary between the two, and column headers are repeated in the third table. The fourth table includes some, but not all, of the columns in tables 2 and 3, while also including multiple subcategories in a 3-

column layout – deviating in design from the prior two tables where those subcategories were styled as rows and not columns.

Figure 7. Variations in tables from Oregon 2015-16 report



SCHOOL PROFILE

ENROLLMENT 2015-16 854

MEDIAN CLASS SIZE	School	Oregon
English Language Arts	22.0	24.0
Mathematics	22.0	24.0
Science	25.0	26.0
Social Studies	22.0	27.0
Self-Contained	--	--

SELECTED DEMOGRAPHICS

Economically Disadvantaged	60%
Students with Disabilities	14%
Ever English Learner	8%
Different Languages Spoken	11
Regular Attenders	68.5%
Mobile Students	23.7%

PROGRESS ARE STUDENTS MAKING ADEQUATE GAINS OVER TIME?

SCHOOL PERFORMANCE

The Smarter Balanced and alternate assessments have four performance levels where levels 3 and 4 are meeting the standard for school and district accountability.

See report cards

Performance of students enrolled in the school for a full academic year

Did at least 95% of the students at this school take required assessments? ☐ Yes ☒ No, Interpret Results with Caution

Participation rate criteria are in place to ensure schools test all eligible students.

	School Performance (%) 2012-13	2013-14	2014-15	School Performance (%) 2015-16	Oregon Performance (%) 2015-16	Like-School Average (%) 2015-16
English Language Arts *(Administered statewide in grades: 3-8, 11)*				*Level 1*	*Level 2*	*Levels 3 & 4*
All students in tested grades	*2014-15 was the first operational year of the new English language arts assessment.*		84.4 12.0 3.6	81.1 12.4 6.5	70.0 17.4 12.6	66.9 20.4 12.6
Mathematics *(Administered statewide in grades: 3-8, 11)*				*Level 1*	*Level 2*	*Levels 3 & 4*

OUTCOMES WHAT ARE STUDENTS ACHIEVING IN HIGH SCHOOL?

		School Performance (%) 2012-13	2013-14	2014-15	School Performance (%) 2015-16	Oregon Performance (%) 2015-16	Like-School Average (%) 2015-16
FRESHMEN ON-TRACK TO GRADUATE	Students who earned 25% of the credits required for a regular diploma by the end of their freshman year.						
	Freshmen on track to graduate within 4 years	--	69.5	82.8	81.3	83.5	82.9

Note: Graduation methodology changed in 2013-14.

		School Performance (%) 2011-12	2012-13	2013-14	School Performance (%) 2014-15	Oregon Performance (%) 2014-15	Like-School Average (%) 2014-15
GRADUATION RATE	Students earning a standard diploma within four years of entering high school.						
	Overall graduation rate	70.7	61.8	71.4	78.3	73.8	77.0
COMPLETION RATE	Students earning a regular, modified, extended, or adult high school diploma or completing a GED within five years of entering high school.						
	Overall completion rate	88.1	88.2	87.4	80.5	81.6	83.9

OUTCOMES FOR KEY STUDENT GROUPS AT THIS SCHOOL COMPARED TO THE SAME GROUPS STATEWIDE

STUDENT GROUP OUTCOMES

	School Performance (%)	Oregon Performance (%)	Like-School Average (%)
Economically Disadvantaged			
On Track	77.4	76.1	78.3
Graduation	72.0	66.4	72.9
Completion	78.8	76.2	80.4
Dropout	3.1	4.3	3.0
English Learners			
On Track	87.5	79.8	86.7
Graduation	64.3	66.9	74.9
Completion	64.3	73.4	80.4
Dropout	3.0	5.0	2.3
American Indian/Alaska Native			
On Track	*	73.3	83.0
Graduation	75.0	55.0	68.8
Completion	100.0	67.4	75.2
Dropout	5.6	8.6	5.2
Asian			
On Track	>95	>95	88.9
Graduation	100.0	87.5	75.9
Completion	100.0	91.2	85.5
Dropout	0.0	1.3	0.7
Native Hawaiian/Pacific Islander			
On Track	*	79.9	91.3
Graduation	--	63.2	68.8
Completion	--	76.6	85.7
Dropout	0.0	5.9	1.6
White			
On Track	82.4	85.1	82.9
Graduation	81.5	76.0	77.7
Completion	84.7	83.8	84.9
Dropout	3.0	3.9	2.9

Although most reports were similarly inconsistent in their designs, Texas' reports were a notable outlier. Across all three reports sampled (i.e., 2004-5, 2010-11, and 2015-16), Texas' reports maintained nearly identical styling and format. Like Oregon's 2015-16 report card, Texas' 2010-11 report card is only four pages in length. Unlike Oregon, however, the Texas report is entirely monochrome, with all data reported in spartan computer-generated tables (see Figure 8). Moreover, every table in the report is identical. The report headers, the rows labels, the column headers, and the data themselves are all presented consistently. The font family and font sizes are uniform. The spacing between rows and columns is static.

Figure 8. Tables from 2009-10 Texas report card

T E X A S E D U C A T I O N A G E N C Y — Page 1

School Name: PLANO SR H S
School Number: 043910001
District Name: PLANO ISD

2009-10 School Report Card

School Enrollment: 2,664
Grade Span: 11 - 12
School Type: Secondary

		State Average	District Average	School Group Median	School (All Students)	African American	Hispanic	White	Native American	Asian/ Pac.Is.	Econ. Disadv.
TAKS Met 2010 Standard (Sum of All Grades Tested) (Standard Accountability Indicator)											
Reading/ELA	2010	90%	96%	96%	99%	98%	97%	99%	> 99%	> 99%	97%
	2009	88%	95%	96%	98%	96%	94%	99%	*	98%	94%
Mathematics	2010	84%	94%	86%	97%	83%	90%	98%	> 99%	99%	86%
	2009	80%	92%	85%	96%	83%	82%	98%	*	98%	84%
Science	2010	83%	93%	91%	98%	93%	95%	99%	> 99%	99%	93%
	2009	78%	90%	88%	97%	91%	89%	99%	*	97%	86%
Soc Studies	2010	95%	99%	98%	> 99%	> 99%	99%	> 99%	> 99%	> 99%	99%
	2009	93%	98%	98%	> 99%	98%	99%	> 99%	*	> 99%	98%
All Tests	2010	77%	90%	82%	95%	81%	90%	97%	> 99%	98%	84%
	2009	72%	87%	81%	94%	81%	78%	96%	*	96%	79%

TEXAS EDUCATION AGENCY
2009-10 School Report Card

Page 2

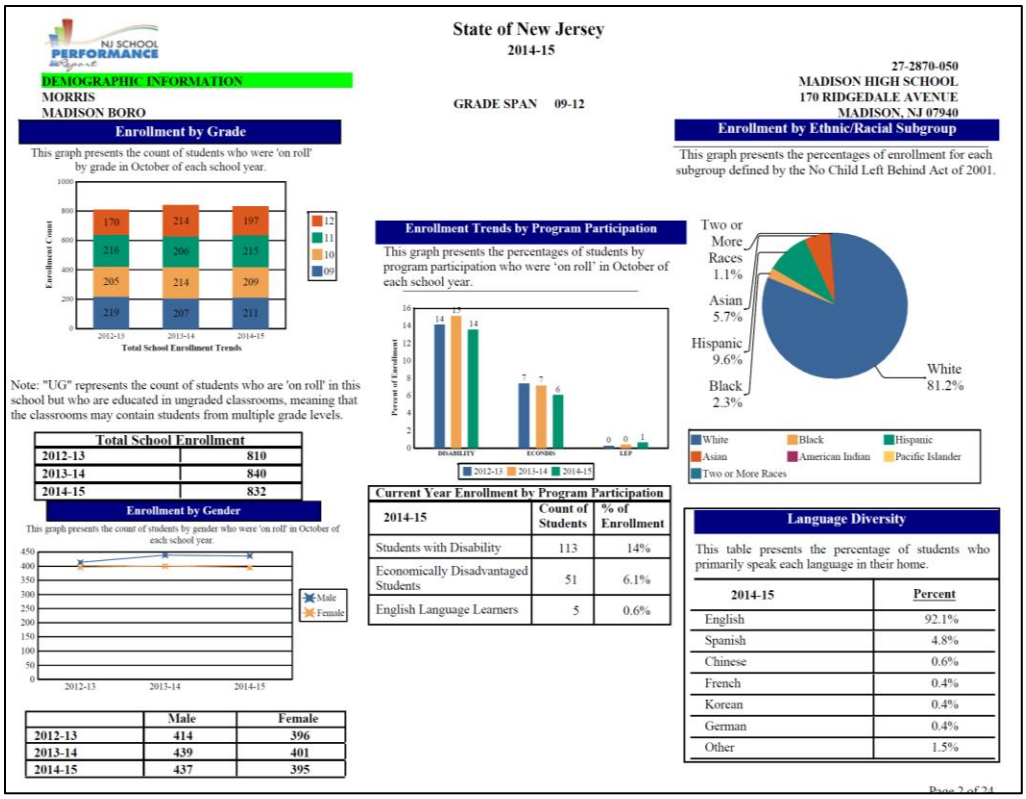
School Name: PLANO SR H S
School Number: 043910001
District Name: PLANO ISD

School Enrollment: 2,664
Grade Span: 11 - 12
School Type: Secondary

	State Average	District Average	School Group Median	School (All Students)	African American	Hispanic	White	Native American	Asian/ Pac.Is.	Econ. Disadv.
Annual Dropout Rate (Gr 9-12)										
2008-09	2.9%	0.7%	0.5%	1.1%	2.1%	2.8%	1.0%	0.0%	0.4%	2.4%
2007-08	3.2%	0.6%	0.6%	0.8%	1.6%	3.7%	0.5%	0.0%	0.0%	3.0%
Recommended HS and Distinguished Achievement Program Graduates										
Class of 2009	82.5%	82.2%	85.6%	83.5%	61.5%	61.9%	85.0%	*	93.9%	55.6%
Class of 2008	81.4%	82.6%	87.5%	82.1%	62.5%	58.5%	83.2%	100.0%	92.4%	56.6%

Shifting our focus from the use of data tables to the use of charts and graphs, the sample reports displayed inconsistent and occasionally inaccurate use of design elements. The most extreme example of these issues arose in New Jersey's 2014-15 accountability reports. In the New Jersey reports, a full page is dedicated to reporting demographic information about the school, including the percent of students enrolled in each grade, the total number of students by race and gender, and the change in demographic representation over time (see Figure 9). Although the content on the page is consistent (i.e., measures of student demographics) the visual representations are not. Each piece of content is displayed in a unique manner, ranging from stacked bars, grouped bars, pie charts, line charts, and data tables. Almost none of the major sections of the report are aligned with one another, or to a common grid, while the styles and formats of design components vary from one instance to the next.

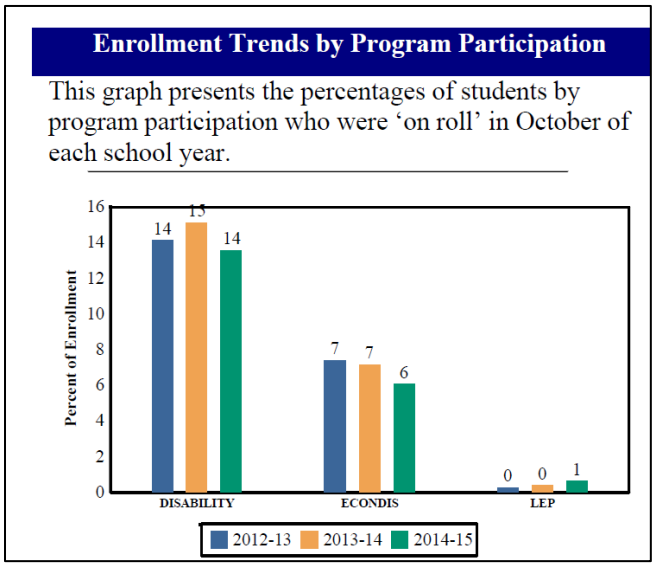
Figure 9. Demographic information in New Jersey's 2014-15 reports



Moreover, this excerpt includes one of the visible errors in reporting identified in the research sample. Figure 10 includes a magnified view of Figure 9, focusing solely on the center charting element of "Enrollment Trends by Program Participation". In this detailed view, one can see a bar chart showing how the enrollment levels of various student groups at this school have changed over time. Although data labels are provided, the actual student groups are referred to with abbreviations of technical school accountability terms, which are not clarified anywhere in the report (i.e., "disability" refers to students with disabilities, "econdis" refers to students who are economically disadvantaged, and "lep" refers to students with limited English proficiency). Looking at the bars for students with limited English proficiency, there is a discrepancy between the size of the bars and the data labels presented. The data labels suggest that this school had

no students with limited English proficiency for 2012-13 or 2013-14, while the bars suggest that enrollment for this group gradually increased year over year.[1]

Figure 10. Enlarged excerpt of "Enrollment Trends by Program Participation"



Explanatory Text

Pre-identified content. Looking across all data visualizations for pre-identified content areas, data were collected indicating whether or not each visualization was accompanied by adjacent explanatory information. For this analysis, explanatory text was defined as any text comprising at least one complete sentence in length; text was considered adjacent to a display if no other data displays and/or textual content appeared between the text and the respective data display. Across all reported metrics in all sample states, over half included explanatory text adjacent to the data visualization; however, the distribution across states varied widely: e.g., Maryland, New Jersey, and

[1] Most likely, the actual enrollment for students with limited English proficiency in 2012-13 and 2013-14 was less than .5%, as evidenced by the varied height of the bars; however, because the data labels are presented as integers, these values were likely rounded down to 0% in the display.

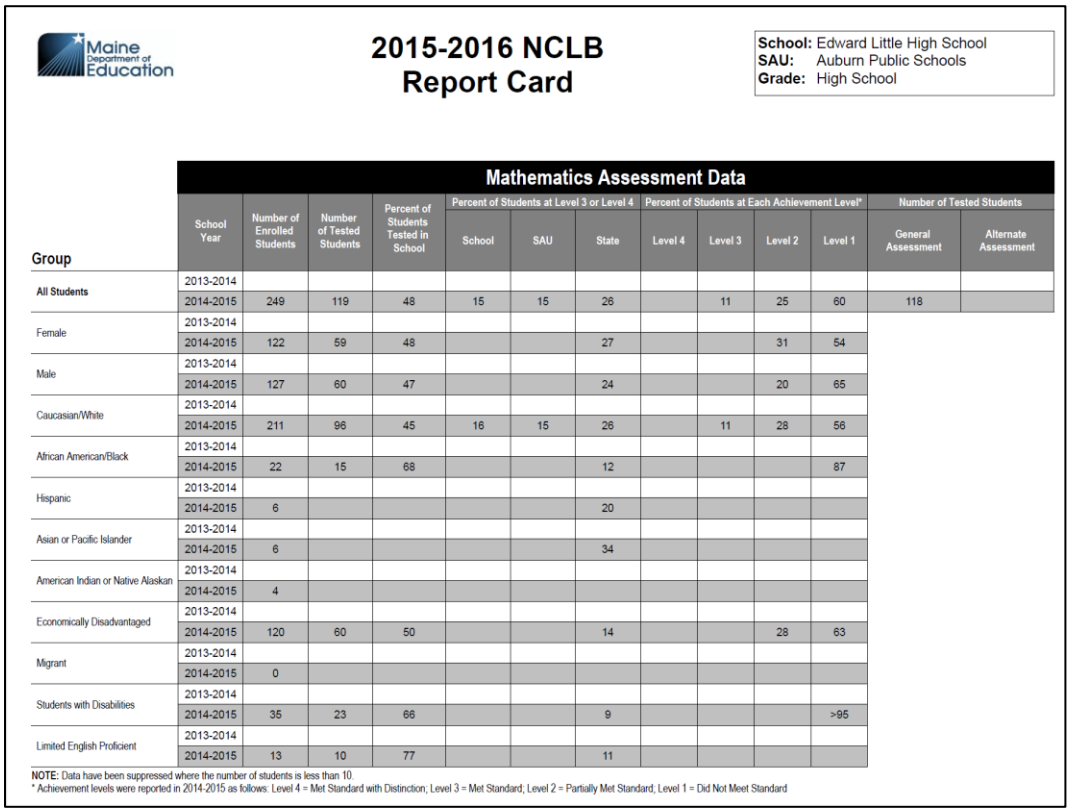
New Mexico included explanatory text for all included metrics; Maine and North Dakota included none. In those instances where explanatory text was included, the average grade level of text was approximately 12.3 on the Flesch-Kincaid scale and 15.3 on the Coleman-Liau scale.

Holistic report review. Looking beyond these pre-identified areas of content, there was significant variation in states' use of explanatory text. Two of the nine states in this sample (i.e., Maine, North Dakota) included no explanatory text throughout the report, but instead included a dedicated page of key terms and metric descriptions that was physically separated from the corresponding data visualizations. In contrast, two states (i.e. New Jersey, New Mexico) included explanatory text for nearly every data visualization included on the report, placing that text immediately adjacent to the visualizations themselves. In contrast, the majority of states (i.e., Maryland, North Carolina, Oregon, Texas, Wisconsin) were inconsistent in their use of explanatory text, providing it for some measures, but not others. For this additional text, reading levels were relatively consistent with the text accompanying the pre-identified measures discussed above, with an average of 13.9 on the Flesch-Kincaid scale and 14.7 on the Coleman-Liau scale.

Examples. Maine's 2015-16 school accountability report illustrates what data visualizations *without* explanatory text often look like. Figure 11 provides a full-page excerpt from the report, detailing student performance in mathematics on the end-of-year Maine Education Assessment (MEA). This visualization appears on the sixth page of the seven page report, yet all key information about this visualization is contained on the very first page of the report in a half-page letter from Acting Commissioner of Education William H. Beardsley. In this letter, several key aspects of the accountability report are described, including a section detailing MEA data. This explanatory text describes the

name of the assessment, the key subject areas assessed, the grade levels in which the test was administered, and the date of administration. Looking back to page 6 of the report, none of this explanatory information is referenced, with the exception of the subject assessed (i.e., mathematics) and the year of administration (i.e., 2014-15). Readers are given no clues as to what assessment results are displayed or for which grade levels, nor are they given explanations of the specific metrics included (e.g., descriptions of what "proficient" means, of what "performance targets" are). Taking this same pattern to the extreme, the North Dakota reports also include an introductory page of explanatory text which applies report-wide (i.e., page 3 of the report); however, this single explanatory page supports 46 additional pages of assessment results.

Figure 11. MEA results from 2015-16 Maine accountability report



Maine Department of Education

2015-2016 NCLB Report Card

School: Edward Little High School
SAU: Auburn Public Schools
Grade: High School

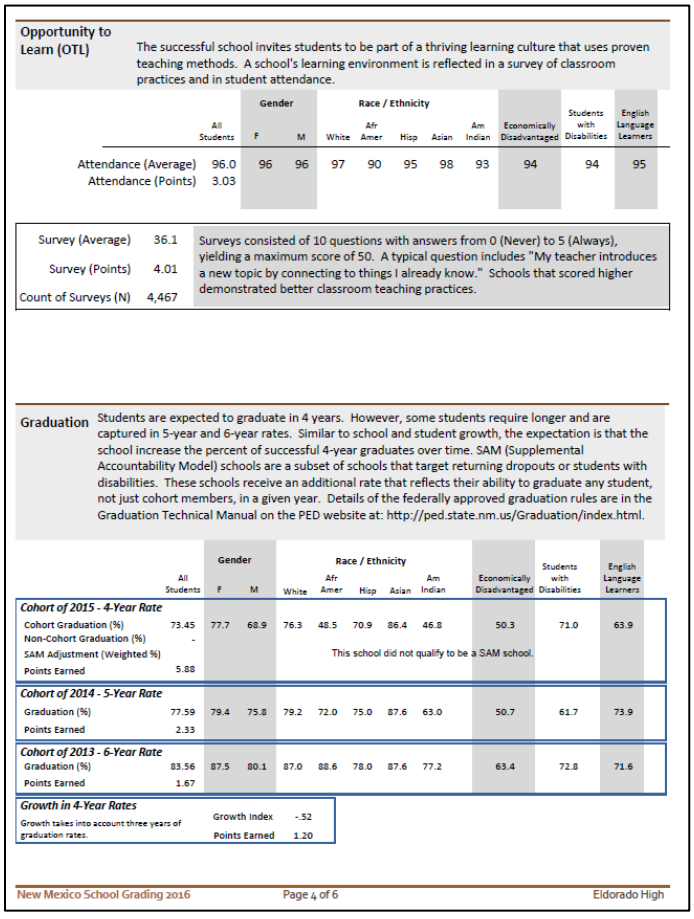
Mathematics Assessment Data

Group	School Year	Number of Enrolled Students	Number of Tested Students	Percent of Students Tested in School	Percent of Students at Level 3 or Level 4			Percent of Students at Each Achievement Level*				Number of Tested Students	
					School	SAU	State	Level 4	Level 3	Level 2	Level 1	General Assessment	Alternate Assessment
All Students	2013-2014												
	2014-2015	249	119	48	15	15	26		11	25	60	118	
Female	2013-2014												
	2014-2015	122	59	48			27			31	54		
Male	2013-2014												
	2014-2015	127	60	47			24			20	65		
Caucasian/White	2013-2014												
	2014-2015	211	96	45	16	15	26		11	28	56		
African American/Black	2013-2014												
	2014-2015	22	15	68			12				87		
Hispanic	2013-2014												
	2014-2015	6					20						
Asian or Pacific Islander	2013-2014												
	2014-2015	6					34						
American Indian or Native Alaskan	2013-2014												
	2014-2015	4											
Economically Disadvantaged	2013-2014												
	2014-2015	120	60	50			14			28	63		
Migrant	2013-2014												
	2014-2015	0											
Students with Disabilities	2013-2014												
	2014-2015	35	23	66			9				>95		
Limited English Proficient	2013-2014												
	2014-2015	13	10	77			11						

NOTE: Data have been suppressed where the number of students is less than 10.
* Achievement levels were reported in 2014-2015 as follows: Level 4 = Met Standard with Distinction; Level 3 = Met Standard; Level 2 = Partially Met Standard; Level 1 = Did Not Meet Standard

In contrast, the New Mexico accountability report provides explanatory text adjacent to data visualizations throughout the entirety of the document. Figure 12 below is an excerpt from New Mexico's 2015-16 report, detailing a handful of measures, including attendance rates, student survey results, and graduation rates. These measures are organized into two distinct groups of information (i.e. Opportunity to Learn, Graduation) with clear explanatory text, detailing what each set of measures are intended to show the reader. Additionally, some measures, like the survey results, are given additional context via adjacent explanatory text, which provides details about the survey design itself.

Figure 12. Explanatory text on New Mexico's 2015-16 accountability report



Opportunity to Learn (OTL)

The successful school invites students to be part of a thriving learning culture that uses proven teaching methods. A school's learning environment is reflected in a survey of classroom practices and in student attendance.

	All Students	Gender F	Gender M	Race / Ethnicity White	Afr Amer	Hisp	Asian	Am Indian	Economically Disadvantaged	Students with Disabilities	English Language Learners
Attendance (Average)	96.0	96	96	97	90	95	98	93	94	94	95
Attendance (Points)	3.03										

Survey (Average)	36.1
Survey (Points)	4.01
Count of Surveys (N)	4,467

Surveys consisted of 10 questions with answers from 0 (Never) to 5 (Always), yielding a maximum score of 50. A typical question includes "My teacher introduces a new topic by connecting to things I already know." Schools that scored higher demonstrated better classroom teaching practices.

Graduation

Students are expected to graduate in 4 years. However, some students require longer and are captured in 5-year and 6-year rates. Similar to school and student growth, the expectation is that the school increase the percent of successful 4-year graduates over time. SAM (Supplemental Accountability Model) schools are a subset of schools that target returning dropouts or students with disabilities. These schools receive an additional rate that reflects their ability to graduate any student, not just cohort members, in a given year. Details of the federally approved graduation rules are in the Graduation Technical Manual on the PED website at: http://ped.state.nm.us/Graduation/index.html.

	All Students	Gender F	Gender M	Race / Ethnicity White	Afr Amer	Hisp	Asian	Am Indian	Economically Disadvantaged	Students with Disabilities	English Language Learners
Cohort of 2015 - 4-Year Rate											
Cohort Graduation (%)	73.45	77.7	68.9	76.3	48.5	70.9	86.4	46.8	50.3	71.0	63.9
Non-Cohort Graduation (%)	-										
SAM Adjustment (Weighted %)				This school did not qualify to be a SAM school.							
Points Earned	5.88										
Cohort of 2014 - 5-Year Rate											
Graduation (%)	77.59	79.4	75.8	79.2	72.0	75.0	87.6	63.0	50.7	61.7	73.9
Points Earned	2.33										
Cohort of 2013 - 6-Year Rate											
Graduation (%)	83.56	87.5	80.1	87.0	88.6	78.0	87.6	77.2	63.4	72.8	71.6
Points Earned	1.67										

Growth in 4-Year Rates		
Growth takes into account three years of graduation rates.	Growth Index	-.52
	Points Earned	1.20

Validity Sample

Overall, these results were found to be largely consistent in the second-round validity sample. With regard to content, all three states (Missouri, Mississippi, and South Dakota) included graduation rates and assessment scores in their accountability reports, making those two measures the only measures consistently reported in all states. Other measures were inconsistently included in the validity sample, matching the pattern found in the first sample (e.g., limited inclusion of achievement gap closure and college-going; more frequent inclusion of demographic data). The one notable difference between the validity sample and the original sample was the inclusion of AMO/AYP measures. While no states in the original sample included this information, two states in (Missouri and South Dakota) included it in their reports.

Similarly, looking at the use of tables and charts, there was consistency between the two samples. Across the three states in the validity sample, every single pre-defined metric was represented using a data table, with the sole exception of student demographic information in South Dakota, which was presented as a pie chart with no data labels and no accompanying data table. In contrast to the frequent use of data tables, these states included very few charts and graphs. Neither Missouri nor Mississippi include any charts in their reports, regardless of content reported. South Dakota diverges from this trend, including a single bar chart for each metric reported; however, the chart only shows information for "all students", while data tables below report data for multiple additional student groups (e.g., race/ethnicity, gender, socioeconomic status). As a result, South Dakota has the most charts per page of any state, in either the original sample or the validity sample; however, these charts include less than 1/10th of the data displayed in data tables.

Looking at micro-design choices, the states in the validity sample were far significantly more consistent than states in the original sample. The reports for Missouri and Mississippi are, like Texas, uniformly designed, with all measures reported in consistently formatted data tables, using uniform color, white space, and font styles and sizes. By contrast, the South Dakota report was least consistent of the three, particularly with regard to the use of white space and the styling of charts and graphs (e.g., the second page of the report includes two pie-charts, one 2-dimensional and the other 3-dimensional, each with different color palettes, legend orientation, point of origin, and title styling).

Finally, with regard to explanatory text, the results in the validity sample largely matched the wide variation seen in the original sample. Both the Missouri and South Dakota reports were largely absent any explanatory text, with each state including it on just one measure in the entire report. Alternatively, the Mississippi report included an explanatory text section for every metric reported. On average, these explanatory sections rated 14.5 on the Flesch-Kincaid scale and 18.0 on the Coleman Liau scale.

Discussion

The findings presented above are necessarily limited by the methods used. By definition, this content analysis examines only the designs as they appear on the page. Consequently, no claims can be made about either the authors' intent when producing these reports, nor about readers' actual interpretations of these reports. Instead, all claims must be limited to statements regarding the reports themselves. Additionally, this research looked at a small sample of all states' reports nationwide. One must be careful extrapolating these findings beyond the sample states considered and applying them directly to other contexts. Furthermore, one must recognize that these reports reflect a specific moment in history. Reports were collected for the 2014-16 school years, after the

enactment of ESSA, but prior to the Trump administration taking office. The content included and the design decisions made are limited to this era of public school accountability. Nevertheless, certain patterns emerge that underscore the importance of attending to the design of accountability reports.

Changing Legislation Leads to Changing Content

First, it is worth mentioning the large variation in content found in these sample reports. Simply put, few states reported on the same content as predicted by the ECS database. For example, based on the ECS data, states in Cluster 1 (MD, ME, TX) were less likely to report attendance; however, both Maryland and Texas included this information. Similarly, states in Cluster 2 (NJ, NM, OR) were more likely to report on data related to college readiness: e.g., AP and IB, ACT/SAT, and college matriculation. However, only New Jersey included college-going data, while Oregon included no data related to AP/IB or SAT/ACT. Because of these differences in content between what the ECS model predicted and what the contemporary reports actually included, we must refrain from making sweeping statements about the reporting practices of groups of states.

However, this variation in content reinforces the idea that accountability legislation has given states more freedom to report on the measures that they care about most. In other words, the variation in content between the ECS database and the current data sample is an indicator of the changing regulatory landscape. Despite being the most up-to-date resource on accountability reporting content, the ECS database primarily reflects reporting practices from the NCLB and NCLB waiver eras. With the shift from NCLB waivers to ESSA, states are given even more freedom to dictate the content of their reports. These research findings suggest that states are using this freedom. Perhaps the

clearest evidence for this shift in policy leading to a shift in states' reporting is AYP/AMO. In the ECS database, over half of all states reported an AYP/AMO measure, which makes sense, given that NCLB required states to calculate and report them. However, in the 2014-16 reports, no state included this information. Yet, despite this variation in reported content, I have attempted to highlight states' varied approaches to representing the exact same content (e.g., assessment results, graduation rates), while also examining how design choices within a state vary over time, and even within a single year's report.

Data Tables as a Mechanism of Regulatory Compliance

Several findings emerge from a close examination of form. First, the sheer dominance of data tables over charts and graphs is worth noting. For the pre-identified content areas, states relied almost exclusively on data tables. While a small few were also reported as charts or graphs, those were the exception, not the rule. Even more illustrative, when looking beyond these commonly reported measures and taking a holistic view, the pattern remains. The two states with the most charts only included them on half of the pages of their reports, while the two states with the fewest charts included none whatsoever. And, yet, this minimal amount of charting was a marked increase from the reports 5 and 10 years prior. This is even more surprising given that one of the very few reporting guidelines provided to states suggests that states include *both* tables and charts in their reports (i.e., "Are the data available in both chart/graph and table format?").

Though this may seem like a somewhat muted finding, the reliance on tables over charts is fascinating. On the one hand, based on current ESSA regulations, one might expect to see the same variation in data displays that one sees in content. After all, while

ESSA provides states with flexibility in the content they include on their reports, the legislation offers even more flexibility in how that content is displayed. The legislation provides no oversight to states regarding the designs themselves. Yet, despite this freedom, states are remarkably consistent in their choice of design.

One interpretation of this finding is that states are approaching accountability report design through a lens of regulatory compliance, rather than one of public dialogue. As mentioned, the literature in data visualization suggests that, relative to other design options, tables are remarkable for their ability to provide a large amount of very precise information in a small amount of space (Tufte, 1997). Because of this, they also make it easier for exploration and investigation of the data (Wainer, 1997). By reporting all measures in data tables, states are able to provide a wealth of information at once to check the regulatory box, ensuring that they have published each and every measure required by the legislation. The use of data tables (and the frequent inclusion of comparison data) allows any stakeholder to answer nearly any ad hoc question they might have.

While this might seem like an equitable approach, it points to a key tension in accountability report design. On the one hand, accountability legislation encourages report designers to take a compliance approach, including all mandated content. On the other hand, accountability reports are intended to provide parents and the public with clear, accessible information about school quality so that parents and the public are able to hold schools accountable. Here we see these two goals at odds. As MacDonald-Ross reminds us, there is no such thing as a neutral design. In prioritizing regulatory compliance via data tables, report designers make it more challenging for readers to make sense of the content included. Although data tables allow audiences to answer nearly any question, they provide audiences little guidance in doing so. Readers are left

to their own devices when searching for which comparisons are most meaningful, for evaluating the relative differences in values over time and across measures. Research suggests that this is non-trivial work, and, work that charts and graphs often accommodate better than data tables (Cleveland & McGill, 1984; Macdonald-Ross, 1977; Schonlau & Peters, 2012).

Lack of Oversight Leads to Inconsistent Design

Another key finding is the inconsistency of design choices throughout the sample reports. Once again, one of the very few guidelines given states encourages states to create reports that are "presented in an understandable and uniform format". Yet, across the sample data, an incredible amount of inconsistency was found not only across reports, but within the reports themselves. Even when looking narrowly at the same types of visualization within a single state's reports, there is variation in the presentation of information – in the size and style of fonts, the alignment of text within tables or on chart elements, with the use of white space, and with the use of color. This inconsistency is worth additional discussion primarily because research suggests that consistency is key to clear data visualizations. On average, audiences more accurately interpret reports that present information in clear and consistent ways, including consistency in the data visualizations used, consistency in the fonts and font sizes, consistency in page layout, consistency in colors, etc. (Gribbons, 1992; Schriver, 1997). Furthermore, in many cases, the reports provided readers with the little to no context for understanding these inconsistent designs. Explanatory text was rarely included alongside tables and charts, and, when it was included, the explanatory text was consistently written at a university level (i.e., grades 12 and above), despite research suggesting that most readers operate at

a much lower level, and despite guidelines suggesting such text should be written closer to an 8th grade level.

These inconsistent designs are more difficult to interpret for several reasons. First, every data visualizations requires decoding and interpreting the component parts; one must identify, for example, the information in a data table by seeking out the table's title, evaluating the commonalities between each element or row (e.g., Is each row in this table a unique grade? A unique subject? Both?), deciphering the relationship between rows and columns, and finally, comparing values across each cell in the table. The process is much the same for charts and graphs. Furthermore, once we learn to decode the first table or chart, we know what to look for in subsequent ones – the units of syntax, the styling of titles, rows, and columns become signposts, helping us re-orient ourselves to new information (Macdonald-Ross, 1977). However, when a report designer changes one of these syntactical elements, they break our mental model.

Although for many readers this act of re-identifying units of syntax may seem trivial, the act of decoding data visualizations requires a solid sense of numeracy and data literacy (Balchin, 1972; Poracsky, Young, & Patton, 1999). Moreover, accountability reports are intended to be read by all parents and families, regardless of their data literacy, and these reports are often read in isolation, without any support or guide from their local schools, districts, or state departments of education. Every inconsistency in design—coupled with the absence of clear explanatory text—is another potential frustration for the reader, increasing the odds that parents and families will either give up on reading through the information, or perhaps worse yet, walk away with an inaccurate picture of what was presented (Hambleton & Slater, 1996).

What makes this finding even more troubling is that it follows directly from the legislative landscape. Like NCLB before it, ESSA provides very little guidance and

absolutely no regulatory oversight on accountability report design. Consequently, in the same way that relaxed oversight on content led to variation in the content that states report, it is no wonder that the lack of oversight on design leads to inconsistent, and occasionally inaccurate, designs. Simply stating that states provide reports in an "understandable and uniform format(s)" is not enough. If anything, one might imagine that the increased autonomy given to states in terms of content reported might compound the problem of designs, with states struggling to fit new content and new designs into old report templates with pre-existing aesthetic and design choices. These findings suggest that current report designs are falling short of their intended purpose. Rather than providing a clear and transparent view into school quality, they are providing an opaque and often incoherent view of school accountability data.

Next Steps

Throughout this study, I have been careful to limit my findings to the documents themselves, avoiding claims that speak to the intent of report authors or the specifics of audiences' engagement with the reports. Nonetheless, both authors and audiences are integral to the conversation. In order to further understand the role that accountability reports play in the broader conversation of public school accountability and public school education, future research must focus on both. Future research is required to identify whether (or perhaps, how) these biases are knowingly created and embedded into the reports, as well as whether the extent to which those embedded biases impact the consumers of these reports. Direct engagement with the consumers of these reports will provide a rich understanding of the politics of (re)production, the intentionality (or lack thereof) of representation, the struggles of interpretation, and the lived experiences of individuals who interact with these reports.

The current research is a necessary first step to support these future engagements with report producers and report consumers. If we want to understand how producers and consumers of accountability reports interact with accountability reports – and therefore, better understand the levers of high stakes accountability in public education – we would be remiss if we did not first have a strong grasp of current reporting practices, as well as the inherent bias in those practices. Armed with this understanding, we can better disentangle the interactions between what authors intend, what is encoded on the page, what is interpreted by audiences, and how audiences act on the their interpretations to enact change.

OPENING THE BLACK BOX OF SCHOOL ACCOUNTABILITY REPORT DESIGN: AN ACTOR-NETWORK THEORY APPROACH

Abstract

School accountability reports serve as a key mechanism in the system of public school accountability. Research in the field of data visualization suggests that the design of these accountability reports impacts how audiences interpret the information they contain. Moreover, prior research suggests that there is a wide range of variation in how states choose to design their accountability reports. This article provides the results of two case studies which examine how accountability reports are designed at the state level. Findings suggests there are multiple tensions which compromise clarity of design, including the content mandates of the legislation and of state departments of education, the absence of parent involvement in design, and the role of technology as a mediator in design implementation.

Introduction

On December 10, 2015 the Every Student Succeeds Act (ESSA) was signed into law, replacing the era of No Child Left Behind (NCLB) with a new legislative mandate aimed at holding public schools accountable for the quality of education provided to their students. Although the new law brought significant changes to the practice of public school accountability, one key provision remained: all states were required to prepare and publicly distribute accountability reports – documents detailing how students are faring academically at each school and district within the state (ESSA, 2015).

Though seemingly perfunctory, this provision mandating accountability reporting is essential to the overall structure of public school accountability. If the public

is to hold its institutions accountable, the public must have a window into its institutions (McGuinn, 2016; Rogers, 2006). The public must know whether, and to what degree, its educational institutions are using tax dollars to provide a high-quality education to all students. Accountability reports serve as that window. These reports provide students and families with a snapshot of how each school and district within the state is fulfilling (or falling short of) its obligation to educate students. In providing this window, accountability reports empower the public to put pressure back onto the school system, holding these institutions accountable for their use of public resources and thereby helping to drive improvement and change.

Yet, despite the incredible importance of these reports, little attention is paid to their design. When accountability reports are discussed, the topic is content (Kane & Staiger, 2001). What information should a state include on these reports? What data will demonstrate that a school has (or has not) properly educated its students? How does one ensure that the information displayed is a fair assessment of each school when there is an extreme variation in the resources provided to schools, as well as in the students served at those schools? These are serious and complicated questions worthy of deliberate consideration.

However, there is another even more fundamental aspect of accountability reporting worth investigating. Regardless of what content a state includes on its accountability reports, how is that content presented? What does the report *look* like? Simply put, looks matter. Decades of research suggest that the presentation of information has a significant, nuanced, and undeniable impact on how audiences interpret what is reported (Cleveland & McGill, 1984; Tufte, 2001; Tukey, 1990). The substance of a report is inherently mediated by its style. The very same information,

when displayed *this way* rather than *that,* can leave audiences with a drastically different understanding.

Complicating matters further, the status quo of report design is incredibly variable. Although accountability legislation has always required states to publish accountability reports, no legislation has put clear requirements on the form that these reports take. The legislation mandates that states produce reports that are "concise, presented in an understandable and uniform format, and accessible to the public" (ESSA, 2015). Beyond that, states are given free reign. There is no formal audit or approval of states' designs by the federal government. Separate from the legislation itself, the Department of Education does publish some guidelines for states embarking on the design process. In these guidelines, states are encouraged to solicit feedback from parents and other "likely consumers of report cards" throughout the design process, "to ensure precise and clear communication of the data," and to "avoid jargon not well known to parents" and to use "graphics and artwork [to] improve readability and maintain user interest" (*Every Student Succeeds Act State and Local Report Cards Non-Regulatory Guidance*, 2017). However, it is the states' responsibility to determine how to best achieve these goals. Given these relatively slim instructions, paired with such a wide regulatory latitude, it is perhaps no surprise that current reporting is extremely inconsistent both across and within states (Moore 2017, forthcoming). Even when controlling for content, states often make very different design decisions when reporting out information about school quality to the public.

Consequently, given the importance of accountability reports to the system of public school accountability, as well as the importance of design choices to the overall impact of these reports, it is worth taking a closer look at these reports themselves. Who, or what, is responsible for the final product? How does each state department of

education come to represent their accountability information in *this* way rather than *that*? As a practitioner-researcher these questions are incredibly relevant to my day-to-day work. Over the past several years, I have worked with numerous teams as they have helped states conceptualize, design, and ultimately distribute accountability reports to schools and districts nationwide. Across these projects, I have had a front-row view as consultants worked with department officials to make sense of the legislative requirements, state's priorities, and stakeholders' needs that govern accountability report design.

This article provides a case study of two such accountability reporting projects. Leveraging Bruno Latour's Actor Network Theory (ANT), this paper examines how the various stakeholders, both embodied and imagined, move from a blank canvas into a fully-realized report design. This process starts, in Latour's words, with artifacts "in the making," looking for key moments and decision-points which lead report designers down one path to the exclusion of all others. Following these moments of conflict, I hope to provide practitioners and researchers with a keen understanding of the key moments of inflection – points in the design process where intervention is likely to have the greatest impact on design outcomes – so that they may more thoughtfully approach design decisions in their own future work. Towards this end, this case study takes up the following questions:

- What tensions arise in the making of public school accountability reports?
- Amidst these tensions, which stakeholders are involved and what strategies do these stakeholders use to resolve moments of tension?

The Role of Reports in Public School Accountability

Before moving into the specifics of these case studies, it is important to understand why accountability reports are integral to the contemporary system of public school accountability in the United States. Accountability, as outlined by ESSA, involves holding schools accountable for providing students with a high-quality education. More specifically, it involves holding schools accountable for student outcomes – for external measures of student learning as measured by end-of-year assessments and other college- and career-ready measures (ESSA, 2015). By making this outcome data public, ESSA empowers students and their families to exert pressure on schools themselves, demanding improvements in quality or choosing to relocate to other neighborhoods and districts that will provide better educational opportunities.

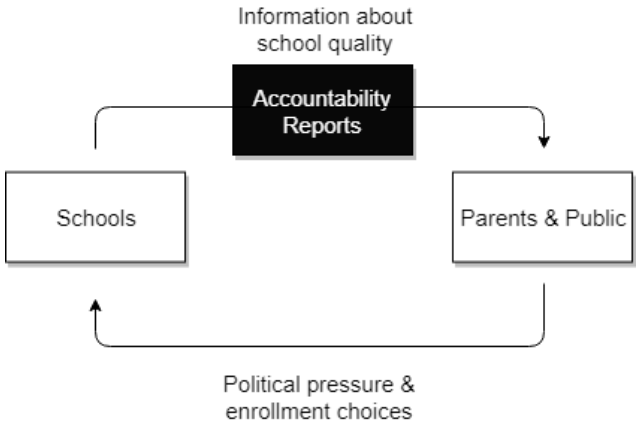
Broadly speaking, the ESSA model of school accountability is inextricably tied to the No Child Left Behind Act of 2002. An extension of President Johnson's Elementary and Secondary Education Act of 1965, NCLB represented a shift in federal education policy. Prior to NCLB, federal oversight of public education in the United States focused primarily on inputs – on the distribution and quality of teachers, on the specifics of curricula enacted in the classroom. Through NCLB, this focus on inputs was replaced with a focus on outcomes – onto measurable markers of student learning (Isaacs, 2003; Wong, 2008).

Put into practice, this shift towards outcomes was really a shift towards assessment. NCLB enacted a sweeping student-level testing mandate for the nation's public school system. Under the law, states were required to implement summative assessments across multiple grades, in both math and English language arts; a concrete deadline was set for all students to reach proficiency on these assessments; and each school was expected to make clear strides towards this deadline in the form of adequate

yearly progress (AYP) (NCLB, 2001). As Kirby writes, "The goal of any program is to bring about desired outcomes. The goal of an evaluation of that program is to determine, through data analysis, whether the program did in fact have an effect on outcomes, and if so, the nature of the effect" (Kirby et al., 2002, p. 142). Through NCLB, student assessment data became the primary source of data for determining a school's impact on student outcomes (Wong, 2008).

Importantly, this model of public school accountability is a model dependent on public feedback (Isaacs, 2003). ESSA requires states to focus on student outcome data, but *also to communicate that information publicly* so that parents and the public may use that information to drive change. Accountability reports matter because accountability reports are the vehicle for feedback. Teachers know "where and how to improve" when they see student scores on standardized assessments – but where do they see these scores? Accountability data helps empower parents "to push for change" – but how do parents access this data? At the end of the day there is a single document, a single report, that must do all of the heavy lifting. The entire system depends on feedback, and this feedback is – for better or worse – embedded into accountability reports. They are the cornerstone.

Figure 1. Role of accountability reports in school accountability



This is why accountability reports matter. These bureaucratic artifacts are integral to an incredibly complicated and incredibly political discourse around the quality of education that our public education system is providing to our nation. Yet, oddly, few have bothered to take a systemic approach to evaluating what they look like and how those design choices come to be.

Theoretical Framework

To understand how these design decisions are made, it helps to take a technological perspective. Although one may not immediately think of accountability reports in technological terms, these reports fit neatly within contemporary definitions of the word. For example, the technological theorist Langdon Winner (1988) provides a three-part definition of technology. For Winner, technologies exist 1) as artifacts, 2) as rules or procedures, and 3) as a series of social processes. Colloquially, one often thinks of the first definition – "technology" refers to artifacts like computers, tablets, phones, etc. However, these artifacts often rely upon a network of rules and procedures. For example, your personal computer depends on an incredibly intricate web of hardware and software, with embedded rules for how to process commands and translate them into your everyday experiences of email, word processing, and internet browsing. The technology is not just the artifact, but also the rules and processes that support how one uses that thing. Moreover, as one uses these artifacts, they quickly become a part one's lived experience. In Winner's words, "as they become woven into the texture of everyday existence, the devices, techniques, and systems we adopt shed their tool-like qualities to become part of our very humanity. In an important sense we become the beings who work on assembly lines, who talk on telephones, who do our figuring on pocket

calculators, who eat processed foods, who clean our homes with powerful chemicals" (1988, p. 12).

Using Winner's definition, accountability reports exist as technologies. At first glance, they are technological artifacts – they are the "object" with which individuals interact. However, beyond artifact, they also represent a series of abstract procedures. Each report represents numerous decisions (or procedures) regarding what content to display and how to display it, what information belongs on the page, and how that information is connected to create a coherent story. Most importantly, these accountability reports also serve a key role within broader social practices. Accountability reports serve as the primary feedback loop within the larger system of public school accountability. The reports connect a broad network of actors – parents, teachers, students, administrators, politicians – serving as the leverage point for the public to pressure institutions of public education to improve.

This perspective, viewing accountability reports as technologies, provides us with a unique entry-point into the discussion of report designs. If one wants to know more about how these reports come to be, about how decisions are made to present information in this way or that, one can start by looking at theories of technological development and technological change. This work will provide a structure for investigating the design and development of accountability reports themselves.

Determinism vs. Social Constructionism

In the early 20th century, Lewis Mumford (1934) described technology as a religious compulsion. Mumford saw capitalist society blindly tethering itself to technological advance, rushing to replace skilled human laborers with cold consistent machines; a shift reflecting a deep-seated desire to master the natural world with

mechanical efficiency. But this mastery came at a cost. “In advancing too swiftly and heedlessly along the line of mechanical improvement,” Mumford writes, “we have failed to assimilate the machine and to co-ordinate it with human capacities and human needs…We have outreached ourselves” (1934, p. 366).

In this exchange, Mumford gets to the heart of what technological theory hopes to understand. What is the relationship between technology and society? To what degree does one influence the other?

At one extreme, theorists argue that the relationship is wholly one-sided, that technology warps and bends society around it. The name for this school of thought is technological determinism. As the name suggests, determinists believe that technology, quite literally, determines the shape of social and political interactions (Smith & Marx, 1994). Its origins trace back to French philosopher Jacques Ellul, who, writing in the mid-20th century, saw technology increasingly independent of human oversight (1964).

For Ellul, the pursuit of technology is a pursuit of efficiency. Tools are created to simplify common tasks, to allow individuals to do more with less. However, as society adopts more and more tools, the drive for efficiency becomes an end in and of itself. Rather than working in service of human needs,

> Technique…pursues no end, professed or unprofessed. It evolves in a purely causal way: the combination of preceding elements furnishes the new technical elements. There is no purpose or plan that is being progressively realized. There is not even a tendency toward human ends. We are dealing with a phenomenon blind to the future (Ellul, 1964, p. 97).

Technology is self-sustaining. While it may have once served the practical goals of its creators – e.g., agriculture, architecture – the nature of technology is towards more, and more efficient, technology.

This efficiency is what ultimately allows technology to determine society around it. In a technological world, efficiency will always be the criteria of success. “When

everything has been measured and calculated mathematically," Ellul writes, "and when, from the practical point of view, the method is manifestly the most efficient of all those hitherto employed…then the technical movement becomes self-directing" (1964, p. 79). In this world, human beings are not agents of change, but rather machines themselves, recording results, and deferring towards whatever option maximizes returns.

This pessimism, however, is only one side of the story. In opposition to the determinists' view that technology drives social structures, other scholars argue that technology and society equally influence one another. These scholars – under the banner of *social constructionism* – suggest that technology never emerges fully grown, but rather develops over time and always in response to various external social pressures. In order to understand why this might be the case, it helps to look at technology "in the making" rather than "ready-made" technology (Latour, 1988). When one looks at existing technology, one sees well-established artifacts and processes; however, when one look backwards, towards the early stages of those artifacts and processes, one often sees that the path taken was anything but efficient and determined.

Two pioneers in this field, W.E. Bijker and Thomas Hughes, demonstrate this point by tracing the development of the modern bicycle. Long before it took its contemporary form, the bicycle came in many varieties – different frames, different materials, different shapes and sizes of tire. Moreover, each design suited a different audience and a different set of needs. For thrill-seekers and racing enthusiasts, the bicycle provided a new outlet for competition and sport. Speed was of the utmost importance; safety was not. As a result, these riders favored a large front wheel, which enabled higher speeds at the cost of a stable ride. For others, the bicycle was merely a means of transportation; safety and a smooth ride were priorities. For these riders, two

equal-sized, air-pressure tires better served their needs (Bijker, Hughes, & Pinch, 1989, pp. 45, 46).

The larger point here is that the development of technology is not set in stone. There was never a platonic ideal of what a bicycle should look like. Rather, the early bicycle – like all nascent technologies – harbored "interpretive flexibility" (Bijker et al., 1989). Importantly for constructionists, this flexibility is a characteristic of technology in the making. Before the bicycle became "the bicycle", stakeholders influenced the development and prioritization of competing design choices (e.g., wheel shape, wheel position). In the same way, constructionists argue, all technologies-in-the-making are fungible. Depending on the stakeholders, and their level of influence, one might end up with *this* tool rather than *that* tool, *this* feature rather than *that* one. The phrase "interpretive flexibility" represents the notion that design and development are responsive to outside stakeholders. To quote Winner, this "social activity [or social construction] is an ongoing process of world-making" (1988, p. 17). When social groups transform technologies like the bicycle or automobile, those social groups also transform the lives of people who use bicycles and automobiles. Had social groups acted differently, different realities might emerge.

Actor Network Theory

Actor Network Theory (ANT) attempts to understand these complexities of individuals, societies, and artifacts in relational terms (Latour, 2007). Relational, in this sense, requires some clarification. The "networks" of actor-network theory are not technological or structural networks (e.g., a broadband network, a freight rail network), nor are they social networks (e.g., individual relationships). Instead, the networks of

ANT are the "fibrous, thread-like, wiry, stringy, ropy, capillary character" of interactions between both individuals and artifacts alike (Latour, 1996, p. 370).

Practically speaking, ANT provides less of a circumscribed theory and more of a loose methodology. In Latour's words, ANT is a way of looking (2007). As researchers attempt to understand technology and society, ANT encourages them to look not at large social structures, but rather at the very minute interactions of individuals and objects in everyday life. It is in these moments, ANT suggests, that one finds the most interesting work being done. It is in these moments that one can see how actors actually *do the work* of constructing everyday life.

To elaborate on this idea, Latour uses the concept of the "black box". In engineering terms, a "black box" is any technology or process that can be understood in terms of inputs and outputs. For a very literal example, one can imagine a vending machine as a black box. It takes input (i.e., money) and returns an output (i.e., food or drink). What *precisely* happens in between is somewhat less clear. Though the general concept is straightforward (the machine evaluates how much money was put in, compares that amount to the price of the item requested, and then returns the item – or not), the specific inner-workings are fairly opaque. What software is used? What mechanism evaluates how much money was given? This obfuscation between input and output is part of what defines a black box.

ANT is an attempt to open up black boxes. To do this, the ANT approach starts by looking for those moments in which the black box was not yet fully-formed. An ANT approach takes us back to the drawing board (sometimes quite literally), to the specific conversations, arguments, and critiques that led to a black box which does *this* and not *that*. Sticking with the vending machine example, an ANT approach would look back to the initial design process. Who was involved in the specific choices made – who were the

actors involved? What happened when actors disagreed over how the black box should operate? How much of the black box was dictated by the available hardware and software? This last question is especially critical to an ANT approach because ANT treats both human and non-human actors equally. Both are critical actors in the development of the "black box" (Latour, 1988).

As researchers trace the winding wiry, stringy, ropy interactions of human and non-human actors, as they open the black box, what they find is something of a battleground. The social construction of technology becomes a political contest, with each of the actors struggling to recruit allies, to gather momentum around their vision. Someone from the sales department pushes for a vending machine that accepts credit cards in an attempt to boost sales; someone in marketing pushes back, claiming that a redesigned display case, one with brighter colors, would increase sales far more than a price hike; an engineer quips that the company could boost sales by improving the universal lock on each machine, reducing the burden on stock workers who have to fill 50 machines a day, and thereby reducing the number of sales lost to empty machines. Each of these are attempts at *enrollment,* the ANT term for mobilizing others towards a particular understanding of what the black box is and what it is not (Callon, 1984).

ANT suggests that these moments of enrollment, many of which often seem rather banal and benign, *are* the moments of social construction. The development of technology is not the result of larger social forces, but rather the end-result of individuals and artifacts constantly waging rhetorical wars, attempting to weave diffuse networks of allies in support of their cause. Simply by limiting one's focus to these actors and enrollments, one begins to uncover "the practical means [used] to keep ties in place, the ingenuity constantly invested in enrolling other sources of ties, and the cost to be paid for the extension of any interaction" (Latour, 2007, p. 66). ANT attempts to escape the

fog of social construction by pointing to the very concrete and lived moments in which "black boxes" come to be.

By looking at accountability reports as technological artifacts – as "black boxes" of their own – ANT provides a guide for better understanding how the ideation, development, and design choices happen within state departments of education. Like Bijker and Hughes' investigation of the bicycle, an ANT approach would tell us to look at reports "in the making". In other words, what actors are involved? How do those actors enroll others? Who is mobilized? And towards which potential outcomes? In Latour's words,

> We are never confronted with science, technology, and society, but with a gamut of weaker and stronger associations; … scientists and engineers speak in the name of new allies that they have shaped and enrolled; representatives among other representatives, they add these unexpected resources to tip the balance of force in their favor (1988, p. 259).

For Latour, the social scientist researcher is part war-correspondent part network-engineer, mapping the highly technical and often elaborate ways in which individuals mobilize resources in their support and levy these resources against other competing views of what could be.

Method

Context

From this vantage point, I spent two years conducting case study research on two separate state departments of education as they worked to conceptualize, design, and produce their school- and district-level accountability reports. Cresswell defines case study research as "a qualitative approach in which the investigator explores a bounded system or multiple bounded systems over time, through detailed in-depth data collection involving multiple sources of information and reports a case description and case-based

themes" (1997, p. 73). My bounded case focused primarily on a for-profit consulting company, Jaxon, which has contracted with states to provide content expertise, design support, and technical assistance to the states themselves. For both design projects, I was embedded with the Jaxon team from the signing of the initial contract through to final delivery and publication of the reports themselves.

Carton. In 2015, the Carton Department of Education (CDOE) released a request for proposals (RFP) for the design and development of a publicly accessible web application allowing parents and families to search for, access, and download PDF versions of school and district accountability reports. Jaxon was one of many organizations nationwide to submit an official response to the RFP and, eventually, was awarded a contract by CDOE to complete the work. The engagement ran from July through December of that year. The work was broadly split between the project team at Jaxon and the project team at CDOE. The Jaxon team consisted of three people: myself, the project manager, Matt, the lead designer, and Hillary, the junior designer.

Throughout the course of the project, Jaxon was responsible for several work products, including project requirement documentation, draft and final designs, building and automating PDF production, and, finally, a live website providing Carton residents with access to all published reports. Project managers from Jaxon and CDOE met bi-weekly by phone to discuss the current project status and to review work products throughout the course of the project. Within Jaxon, the project team also scheduled a weekly project meeting; however, Jaxon team members would discuss the project in person, by email, and via web-conferencing on a daily basis. All project materials were shared across teams using an online project management website. Major deliverables and all major project decisions were documented by email. Design documents and PDF

production were facilitated by Adobe InDesign, while the public-facing website was built using custom code.

Sydney. In 2017, the Sydney Department of Education (SDOE) hired Jaxon to help design their new, public-facing website for sharing ESSA accountability reports. Unlike the Carton project, the Sydney project was limited to designs only – the goal of the project was for Jaxon to produce draft designs and final design documents, which would then be turned over to the SDOE technology department to translate into a working website. Jaxon would not provide any technological support or software development resources to the SDOE. At Jaxon, the project team consisted of two individuals, Doug, the project manager, and Matt, the lead designer.

Unlike Jaxon's work with Carton, the Jaxon and Sydney teams did not meet on a regular cadence; instead meetings were scheduled to coincide with major project milestones and with the delivery of key work products. However, like the Carton project, Jaxon's internal team discussed the project on a daily basis, primarily through in-person meetings, web-conferences, and via an internal messaging service (i.e., Slack). Because Jaxon was hired to provide Sydney with guiding designs (rather than implementing those designs), the primary work products were draft design PDFs, draft interactive designs, and a final interactive design. All interactive designs were created using a publicly available proprietary software called InVision. These design documents, both static and interactive, were shared between Jaxon and SDOE via an online project management site.

Researcher Positionality. Throughout the Carton and Sydney projects, I served as both practitioner and researcher. More specifically, while conducting the case research, I was employed by Jaxon and worked with both the Carton and Sydney project teams. For the Carton project, I served as project manager, facilitating project meetings

with the CDOE project team and managing the internal Jaxon team as they worked to design and deliver accountability reports. For the Sydney project, I served as an executive sponsor. In this role, I provided feedback on initial designs and conducted client-satisfaction reviews with the SDOE project lead.

Undoubtedly, my role at Jaxon directly impacted the research presented below. Because of my position within the Jaxon team, I was given unfettered access to all project materials, design documents, and memorialized conversations (e.g., emails, meeting minutes, online conversations). Moreover, my intimate knowledge of events, coupled with my participation in the projects themselves, allowed me to ask more pointed questions about the design process and the resolution of design tensions based on my unique experience with the participants and with other Jaxon projects outside the scope of this work. At the same time, however, my position necessarily biases both my view of the data, as well as participants' responses during interview questions. Though the direction of these biases is difficult to unpack, care was taken to mitigate them by introducing multiple data sources beyond participant interviews and by coding all data as methodically and independently as possible.

Data Sources

Carton. Jaxon's engagement with CDOE transpired from July 2015 through December 2015. Over this period, the main sources of data collected were project communications and project artifacts. In particular, I collected over 208 distinct email chains related to the CDOE project. Of these emails, just under half (86) involved direct communication between the Jaxon and CDOE teams, while the remainder (122) included internal communications between Jaxon team members. As part of this email

correspondence, I had access to summary notes from 10 project meetings between Jaxon and the CDOE team.

In addition to this correspondence data, I also collected several artifacts. Artifact data emerged primarily from a web-based project site maintained by Jaxon. This site served as a repository for the project timeline, business requirements documentation, and a running issue log for final PDF production. Additionally, the site maintained a database of design artifacts, including 18 discrete design documents, together representing every design iteration from the initial sketches and wireframes of the accountability reports, to interim drafts shared with the client, and to the final published designs.

While all of these data were collected throughout the project, interview data was collected after the project had closed. Interviews were conducted with both the lead designer and the junior designer, each of whom worked on the project from start to finish. Each participant took part in two one-hour interviews, reflecting on the project overall and their recollection of specific design decisions therein, with some interim design artifacts provided for reaction.

Sydney. The SDOE project ran from April 2017 through August 2017. As with the Carton case, correspondence data were collected throughout the duration of the project, including 43 email chains. Of these emails, 38 involved both the Jaxon and SDOE teams, while five represented internal Jaxon communications. This lack of internal email communication was due to Jaxon's adoption of an internal messaging service, Slack. In addition to email correspondence, I collected all internal Slack messages related to the Sydney project. Over 75 messages were collected from the initial project kickoff through completion of the project.

Data were also collected from Jaxon's web-based SDOE project site. These data included project timelines, business requirements documentation, and draft data diagrams (i.e., documents detailing the content to be included in the reports and the variation in that content between schools). Like Carton, the Sydney project site included several design artifacts. Ten design drafts were collected, again running the spectrum from initial drafts through to final designs.

After successful delivery of the project, interviews were conducted with the two primary members of the Jaxon team. Both the Jaxon project manager and the Jaxon lead designer sat for two hour-long interviews, following the same format and protocol as the Carton interviews.

Table 1. Data sources at Carton and Sydney

Data Sources	Carton	Sydney
Email	208 chains	48 chains
Meeting summaries	10 documents	6 documents
Chat logs	N/A	78 messages
Project website	Yes	Yes
Design drafts	18 artifacts	10 artifacts
Interviews	2 participants; 2 1-hour interviews each	2 participants; 2 1-hour interviews each

Analysis

With this data in hand, I approached analysis from an ANT perspective, looking at "technology in the making" as the Jaxon, CDOE, and SDOE teams worked to design their accountability reports. In particular, I started by surveying the conversations between project stakeholders, conducting an initial content analysis to uncover the key actors in each project, and the language that those key actors used throughout the project. Next, I looked specifically for key turning points in the design process, noting where and when the design changed from one draft to the next, as well as how these changes were discussed among the various project team communications. Additionally, I coded for moments of tension and disagreement, both in these conversations regarding design changes, as well as in project team members' self-report during post-project interviews. Finally, within these moments of tension, I coded for moments of *enrollment*, looking at how individuals enrolled other actors in support of their unique perspective to resolve tensions and to "close" the black-box of design.

Content Analysis. Once data were collected, the preliminary analytical task was to determine the major actors in the case study and the substantive content of their communications. To do this, I looked exclusively at the communication-based data, including emails, meeting summaries, online chat logs, and interviews. Within each data source, I identified actors by first-hand speakers as well as second-hand references to others; primary actors were classified based on frequency of communication and/or reference. After identifying these actors, the content of communication was coded by reading through all communications and marking broad themes (e.g., enrollment of allies, technology as actor, parents as actors, expressions of authority) then comparing the presence of these themes among actors within each case and across cases, as well as over time.

Turning Points. The next phase of analysis looked for key turning points within the design artifacts themselves – moments in which elements of a design (e.g., the layout of content, the use of specific data representations, and/or charting components) varied from one design draft to the next. To look for these moments, I relied primarily on the sequence of design artifacts themselves. In particular, I collected all design documents available, including internally produced designs that the Jaxon team never circulated with CDOE or SDOE, as well as the design "deliverables" that Jaxon presented to the client as required by their service contracts. With each iteration, I marked all changes, both major and minor, while distinguishing between changes in content and changes in form, as well as changes that involved both content and form. Examples of changes in content include changes to labels and titles (e.g., renaming a section title from "Academic Progress" to "Academic Growth"), while changes to design might include a shift from displaying student enrollment as a table to displaying that data as a pie chart.

Tensions. Using these key turning points as a guide, I then looked for moments of tension in the communication between stakeholders described above. Here the primary data source was participant self-report. After identifying key turning points in design, I asked interview participants to review design drafts both before and after key turning points (e.g., looking at a draft design from September and comparing it to a draft design from October), and reflect on what led to each shift. In these discussions, I coded for any disagreements or tensions between Jaxon team members or between Jaxon and both CDOE and SDOE, trying to uncover whether these turning points represented moments of disagreement between parties or benign progressions from draft to final design.

Enrollment. Finally, with these moments of tension identified, I looked deeper at the communication between stakeholders for *enrollment*, "[looking] at where the

disputing people go and what sorts of new elements they fetch, recruit or seduce in order to convince their colleagues"(Latour, 1988, p. 15). To do this, I narrowed my focus to project communications and interview transcripts directly related to each moment of tension. Within each piece of data, I coded any moment where actors represented other actors, human or non-human, in defense or in opposition to a design choice. In each of these enrollments, I looked for who was speaking on whose behalf, and how each speaker characterized enrolled actors – what beliefs, attributes, and arguments were attributed to one actor by another? By focusing on these moments of enrollment and challenges to authority I was able construct a case analysis that draws on common themes across these moments to better understand how accountability design decisions were ultimately resolved.

Findings

Content Analysis

Overall, the most common stakeholders discussed during each project were the project teams at Jaxon and at the state departments of education. When reviewing project communications (i.e. emails and chat logs) and counting the total number of references to each stakeholder group, over half of all references were to these project teams. For Carton, this included three Jaxon employees (i.e., myself, serving as project manager, Matt, the lead designer, and Hillary, the junior designer), as well as the CDOE project manager, Ashley, and the Chief Accountability Officer, April. Similarly, for Sydney, the main stakeholders included the Jaxon team (i.e., Doug, the project manager, and Matt, the lead designer) as well as the project management team at Sydney.

In addition to the immediate project teams, the next most frequently discussed stakeholders included "parents" and "the public". References to this group made up just

under one third of all references in project emails and chat logs. As early as the initial project kick-off meetings, team members began to carve a substantial role for parents with regard to project direction and decision-making. For example, at Sydney, the SDOE team began the project kickoff meeting with a discussion of parents as key users of the accountability reports. In Doug's words, "[they] specifically identified the parents and the public as a key user group and also defined those parents to be non-technical users who were inexperienced with the data." Similarly, at Carton, the initial requirements gathering defined the end user as everyday parents who were not particularly data-savvy, but who were concerned about their own students' well-being and education.

Along with this discussion of parents, early conversations primarily revolved around discussions of project expectations, including specifics of the final project deliverables, interim project milestones, and project timelines, as well as in terms of project communication, meeting schedules, and methods. Additionally, early conversations also focused heavily on issues of content. Prior to any substantive design discussions, the Jaxon team pressed both SDOE and CDOE to consider what information would be displayed on the reports themselves. These conversations spanned the first several weeks of the project and involved back and forth between both teams to determine what information was required by law, what data were accurately collected and readily available, as well as what information the state departments of education would want displayed on the reports.

When discussing content, participants often used specific terminology to help determine what to include and exclude from the reports. For example, in the SDOE project, Doug described how SDOE team members leveraged the concepts of "school profiles" and "school report cards" to defend their favored content. In their description, the job of the school report card was to pass judgment on schools, providing parents with

clear information about how each school is serving their individual students. The role of the school profile, by contrast, was to provide families with a holistic view of the school, including, but not limited to students' performance and educational outcomes. In establishing these two different models, the SDOE project team was able to successfully argue that, because their project was a school report card and not a school profile, non-academic content should be removed from the accountability reports.

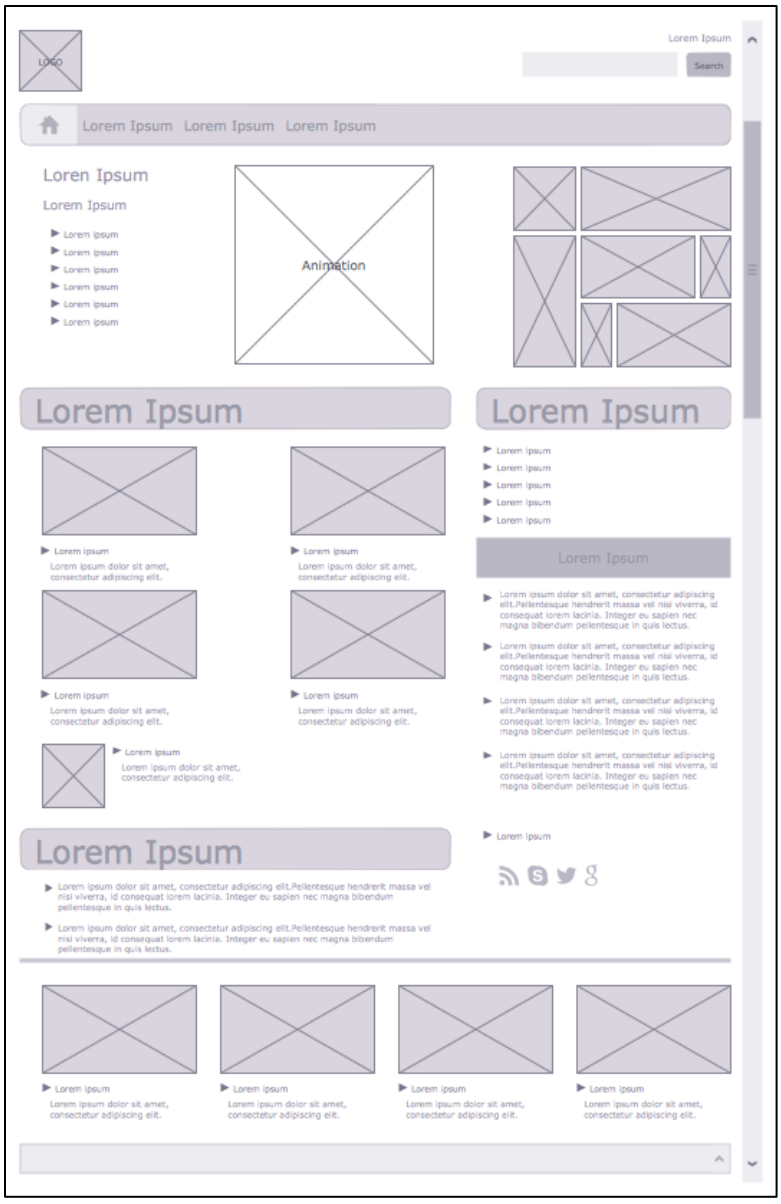
As the projects progressed, conversations split between content and form. Early in both projects, issues of content appeared in over 70% of project communications (i.e. emails and chats), while issues of form appeared in only about 10% of those communications. In later weeks, this balance was entirely reversed, with design decisions making up over 80% of communication and content nearly falling out completely. In the early discussions of content, the Jaxon team primarily deferred to their clients: Jaxon team members viewed the content provided by both Carton and Sydney as fixed and discussed how to best translate these inputs into a design output. As Matt described, "We look to [the client] for content collection and [then] we structure things how we see fit." Similarly, Hillary explained, "During design reviews and when we're structuring wireframes, we're not communicating with the SDOE team." Instead, Jaxon's internal discussions confronted how to best translate the given content into a visual design. These conversations were incredibly detailed and focused, with Jaxon team members communicating back and forth on very specific aspects of the design. For example, after reviewing one of Hillary's CDOE designs, Matt replied with specific font size recommendations on two specific labels (i.e., "Make the Performance Indicators 'Academic Achievement' etc. 9pt. and the Performance Level 'Far Below' etc. 7pt."). Similarly, with the SDOE project, Doug responded to one of Matt's designs with specific recommendations for a table layout (i.e., "What do you think about adding a column for

'proficiency'? and always adding a (sic) 'all students' row at the top so you have some comparison?") In these two examples and across the multitude of communications between Jaxon team members, the content of design conversations maintained this precise tone, with the bulk of conversations addressing relatively specific design choices.

Turning Points

Surveying the design drafts from the Carton and Sydney projects, key turning points in design varied throughout the project lifecycle. In both projects, the first major design changes occurred early into the project, specifically during the *wireframing* portion of design. Wireframes are quick, black-and-white, sketches that frame the broader design. They often do not include any styling (e.g., fonts, icons, images, or data displays), but instead they "block out" the page into discrete chunks, demonstrating how the key information will be organized on the page (see Figure 3 below). Wireframes serve as a very rough draft of the proposed final design, allowing designers to get quick and immediate feedback on their ideas without investing too much time or effort into a more polished and stylized document.

Figure 2. Sample wireframe from www.conceptdraw.com



Because designers use wireframing as a means of problem-solving – of brainstorming potential solutions to design problems – there is often huge variation from one wireframe to the next. In both the CDOE and SDOE projects, early wireframes varied significantly from later wireframes, representing a key turning point for their designs. (Additionally, this transition from wireframe to fully-formatted design

document was universally identified as a turning point during interviews with study participants). Comparing these designs side-by-side, the primary differences are structural. In each wireframe progression, one finds a slightly different approach to the organization of information on the page. For example, in the Carton designs, early wireframes vary in layout of the page, alternating between portrait (i.e. the longer edge of the page on the left and right side) and landscape (i.e. the page turned such that the longer edges are on top and bottom). Similarly, these wireframes also vary in the amount of the page dedicated to each content area. For example, early Carton wireframes equally split the front page of the report, with half of the page dedicated to school information (e.g., principal name, grades served, school mission, student demographics) and the other half dedicated to accountability results (e.g. school performance on key indicators). Later wireframes significantly alter in their approach, eschewing a 50-50 split for a design where school information is de-emphasized, taking up just a quarter of the page, while accountability information is prioritized and given the remaining space.

As the projects progressed, the key turning points in design were more specific, representing changes to one or two design elements, rather than overall structural and organizational considerations. For example, in the Sydney project, one key design change involved different representation of student performance by race and ethnicity from one design iteration to the next. In the earlier design, this race/ethnicity data was styled in the same manner as other student-level information: displayed in a flat data table with precise numbers for each cell of the table. However, in the subsequent design, these numbers were replaced with icons representing performance. Whereas the earlier design might report a specific group's performance as a percentage (e.g., "80% meeting expectation"), this new design used color coding. Green icons represented groups who met expectations, while red groups represented those who did not. Similarly, a key

turning point in the later phase of Carton's reports involved the choice of how to represent school's overall performance rating. Based on a complex rubric, each school was assigned an overall quality rating, equivalent to the conventional A, B, C, D, F grading system in K12 schools. In addition to displaying this rating, April, the Carton Chief Accountability Officer, also wanted to display a visual icon. Design artifacts show Jaxon taking two discrete approaches to solve this problem. In the first, each rating was represented by an icon of a color-coded trophy. The highest rated schools received a green trophy; the lowest rated schools received a red trophy, etc. Alternatively, Jaxon also produced a version of the design where each performance rating was accompanied with an icon of a spaceship being built. Schools who received the lowest rating would receive an icon of the ship's blueprints; schools with an average rating would receive an icon of the ship being built; and, schools with the highest rating would receive an icon of a ship in flight. By providing these alternate designs, color-coded trophies vs. spaceship construction, Jaxon's designers attempted to create two different visual rhetorics – one emphasizing competition and performance (i.e., trophies) and another representing non-judgmental improvement (i.e., taking steps towards flight).

Together, these examples are reflective of the types of turning points found across all design artifacts. Progression through early wireframes demonstrated large-scale shifts in the hierarchy of information and the relative real-estate given to content, while later design shifts were more limited, focusing on specific design choices within a relatively established structure. As each project's final deadline approached, the number of turning points and their relative impact, became smaller and smaller.

Tensions

Across these key turning points, moments of tension were relatively infrequent and muted; however, moments of disagreement did occur. During post-project interviews, participants unanimously identified three distinct types of tension in each project: tensions between Jaxon and their clients, tensions between Jaxon project team members, and tensions with technology. For the former – tensions between Jaxon and their clients – these disagreements played out during client design reviews, meetings in which the Jaxon team deliberately walked clients through one or more design artifacts and solicited specific feedback. Within Jaxon, tensions between project team members occasionally occurred in the same circumstances, during deliberate design reviews; however, many of these tensions also played out in digital communication, via email or online. Tensions with technology were more difficult to locate within project meetings and communication, but often came out during moments of reflection during participant interviews.

In moments of tension between Jaxon and its clients, the resolution often involved Jaxon bowing to client demands. For example, a clear external tension occurred in the Sydney project related to the student race/ethnicity display discussed above. When reviewing the initial Jaxon designs (i.e., the version with data tables rather than the version with color-coded flags), the SDOE team expressed dissatisfaction. In prior years, when SDOE produced their own reports, they often used color-coded flags. Moreover, as they explained, SDOE had received phone calls from parents who voiced appreciation for the flags on their prior year reports. Because parents liked the design, SDOE wanted to maintain it on their new reports. However, despite this argument, the Jaxon team clearly preferred their own approach. Countering SDOE, the Jaxon team argued that the particular icons used would potentially distract parents. In Matt's words,

"I'm not really sure as a parent I would really understand what's happening." Similarly, Doug pressed the argument further, stating that "more minimal design is less likely to overwhelm [parents]," and suggested that parent feedback was potentially misrepresented. In his words, "You remember the one other [phone call] that happened and together in your mind you have this sample of times when people asked about [the design]. What about all the people who didn't call you?" Yet, despite these rebuttals, the SDOE team persisted. After Jaxon pressed back on SDOE during the design review, SDOE asked for the next version of the design to include the flags, and the Jaxon team complied without further argument.

Within Jaxon, tensions between team members were often resolved by a deference to mutually agreed-upon goals. During interviews, Jaxon designers explained that their first goal in any project was to clearly define *users* and *goals*. In both the Carton and Sydney projects, the Jaxon team defined the primary users of accountability reports as parents and families who did not have a technical background and who did not have an expertise in data and analytics. The goal for these users was to quickly and easily understand the quality of service provided by each school, as defined by the state's system of accountability. With this in mind, design decisions were always evaluated based on whether or not they helped these users achieve this goal. For example, as discussed, one key turning point in the CDOE design process involved the amount of space allocated to school information versus accountability information. Because the goal of these reports was to help parents quickly understand school quality, Jaxon chose to prioritize accountability information – and thus the design progressed from one in which school information and accountability information were treated equally, to a design where accountability information dominated the page. This strategy of

scrutinizing competing designs against the proclaimed report users and those users' goals helped to resolve the majority of internal tensions at Jaxon.

Separate from these interpersonal tensions, technology also served as a point of tension for the Jaxon team. For example, during a post-project interview, Hillary described how tensions emerged between Jaxon's designers and the software they used to produce CDOE's reports. Practically speaking, Hillary explained, Jaxon used an Adobe software (i.e., InDesign) for its automated PDF production, while web design was hand-coded by Jaxon's software developers. While working on the cover page of the PDF reports, Hillary noted that the Jaxon team struggled to find the right visualization for student demographics. Reviewing early drafts designs, one sees this struggle on the page, as competing design documents alternate between visualizing the proportion of students within each racial and ethnic group by using a traditional data table, a colorful pie chart, or a series of horizontal bars. During an internal design review, Hillary recalled, she argued that the pie charts and horizontal bars would work best if they were ordered by size. In her words, "We know from experience that a pie is really hard for parents to read and to make it easier, we [could] stack the largest percentage first from the biggest to the smallest" – in other words, the first slice of the pie would represent the largest student group at the school, while the smallest slice would represent the smallest student group. Her reasoning was that parents rarely want precise information about each and every student group, but parents do care about their individual children. She explained that, "parents' mindset is [focused on] their one kid and where do they fit in. If my kid is 2% in this report, I wouldn't weigh this report too heavily." In other words, by showing demographic groups from largest to smallest, Hillary believed parents would be able to more easily see whether the majority of students at the school resembled their own child, and evaluate the report accordingly. However, Hillary immediately countered her own

argument by appealing to tensions between her vision and the capabilities of the InDesign software. InDesign, she said, would not be able to do this ordering. Because each school has a different demographic makeup, each report would have to dynamically re-sort the list of student groups accordingly, something that InDesign would not do. This argument was accepted without challenge, and the team decided to use data tables in the final design.

Enrollment

In resolving these tensions, actors often relied on the ANT strategy of enrollment – appealing to other actors as allies in defense of one's own particular position regarding each design decision. The most prevalent enrollment strategy was an appeal to "parents". Across both the Carton and Sydney projects, both within the Jaxon team and within the state departments of education, stakeholders defended their design decisions by appealing to parent needs and desires. Beyond enrollment of parents, enrollment of technology was also a key strategy in both projects. Technology was leveraged as the arbiter of disputes, with the winning design often, if not always, being the design more easily accommodated by available technology.

Parents. Across both cases, "parents" emerged as a critical actor within the actor-networks of the Jaxon, Carton, and Sydney project teams – despite the absence of actual parents participating in the projects. In both projects, "parents" first emerged during the project requirements gathering phase described above. Once parents were established as key actors, the details of who these actors were and what these actors valued became a rhetorical battleground. Whenever there was a disagreement over a design decision – whether among Jaxon team members or between members of the

Jaxon team and employees at CDOE and SDOE – actors would work to reshape and redefine "parents" in support of their own individual case.

Throughout the project, multiple strategies were used to reposition parents in this way. As mentioned, both projects began with a straightforward, if vague, assertion of parents as users who were not well-versed in data and data visualization (e.g., "These are made for regular people"; "Less tech savvy or data savvy"; "Non-technical users who were inexperienced with data"). Beyond these minimal, initial descriptions, the actual attributes of parents remained relatively unexplored throughout the project lifecycle. Instead, building on this sparse definition, project team members found their own unique ways to divine these actors' needs and values. For example, Matt often used his own past first-hand experiences with parents in other cities and states to shape the parents of Carton and Sydney. When arguing for the use of bar charts over pie charts, or over the maximum page limits of the reports, Matt defended his choices by arguing that those choices were always supported by his personal interviews and focus groups with parents (e.g., "We've talked to a lot of parents"; "A lot of [decisions] are backed by projects we've done"; "We've actually had a good breadth of experience where we can talk to people about [design choices] and point them towards past experience"). In each of these cases, the unstated assumption was that Jaxon's past experience with parents was applicable to Sydney and Carton; that past parents were similar to parents in new locales. Another key strategy involved deferring heavily to "existing research" when making assertions about parents as actors. While working on the Carton project, for example, Hillary described rooting through census data, and even going so far as to pick a specific county and town at random, and "[tried] to imagine a persona that way". In interviews, she described arguing internally with the Jaxon design team, and defending her design choices by deferring to this research (e.g., Hillary described that the census

data provided "good background information to know so you can come up with good arguments [to defend a design]").

Although parents were an assumed actor within both projects, the Jaxon team members expressed an awareness that these assumptions were being made and also an awareness that such assumptions were often spurious. Participants suggested that the disembodied "parents" often became vehicles for voicing personal beliefs and assumptions rather than vehicles for accurately representing parents' needs. For example, in reflecting on the design process, Matt lamented the fact that clients often ignored outside definitions of parents' needs and interests in favor of their own preconceived notions: "they think that their users are different, and they understand them best". Doug echoed this sentiment, adding that, "I think they would say that they were thinking about other folks more, but I think what they really do is think about folks like themselves."

Interestingly, though, neither Matt nor Doug voiced this same skepticism in relationship to themselves or to the Jaxon team. For example, in reflecting on the Sydney project, Matt expressed that, "Whether accurate or not, there are usually some assumptions [by the client] about who the user actually is, but we've had a good breadth of experience where we can talk to people about it and point them towards existing research and [our] past experience. They put a lot of trust in that." In this brief reflection, Matt not only casts doubt on SDOE's ability to accurately speak on behalf of parents (i.e., "whether accurate *or not"*) but also simultaneously establishes himself and Jaxon as the authority on the parents (i.e., "a good breadth of experience", "a lot of trust"). This pairing – a skepticism towards clients' ability to speak for parents, yet a trust in Jaxon's own ability to do the same – was consistent across all participant interviews.

Technology. Another recurring theme throughout both the Carton and Sydney projects was the enrollment of technological actors in support of design decisions and to resolve disputes over competing designs. For Jaxon, this often involved personifying their software products and software code as an actor with its own needs, preferences, and capabilities supportive of certain designs and at odds with others. Unsurprisingly, this personification was most frequently heard by project members tasked with implementing the designs rather than imagining the designs (e.g. producing an actual web report with real data, as opposed to designing a mock report with mock text). Moreover, while this strategy was frequently used by team members at Jaxon, it was not adopted by their clients even when their clients were tasked with implementation work.

One key example of this strategy arose during the Sydney project. For Jaxon, the Sydney project was design-only. In other words, Jaxon was responsible for creating an overall template for Sydney's accountability reports, but SDOE was responsible for taking those designs, implementing them, and producing the final reports in-house. As a result, there was no immediate need for Jaxon to consider whether the design could or could not be produced by Jaxon's preferred tools. Nonetheless, this strategy persisted. As Doug described, the final Sydney design decisions were "the result of Jaxon trying to make designs it could ultimately implement if it were tasked with [that work]". As a result, several design disputes were resolved much like the race/ethnicity dispute was resolved in the Carton project. However, here the defiant actor was not InDesign, but instead Jaxon's own in-house product suite, informally referred to as HiFi.

For Jaxon's clients, HiFi was pitched as a powerful tool to quickly and flexibly create public-facing reports. Clients would put their schools' information into pre-existing data templates, which Jaxon would then feed into HiFi. Almost instantly, the clients' data would be transformed into a highly interactive, user-friendly website ready

for public dissemination. Each of the design choices in the HiFi website reflected Jaxon's best design thinking, based on the firm's own wide-ranging past experiences and client feedback. However, behind this user-facing site, HiFi ultimately existed as a collection of custom software code written by Jaxon's programmers over the course of 18 months. Like all software, this code was written based on certain assumptions and in alignment with certain product requirements. For example, HiFi required client's input data to meet the exact standards of the provided data template. Similarly, HiFi's data displays – although customizable and configurable – could only vary within certain pre-defined boundaries.

Throughout the Sydney project, the Jaxon team used these limitations to defend certain design choices and to challenge others. For example, very early in the project, Jaxon's Chief Product Officer (i.e. the person in charge of HiFi) requested an internal design review of the Sydney designs. At this phase of the project, the only design documents were wireframes detailing informational hierarchy – no data visualization choices had yet been made. Nonetheless, the Chief Product Officer pushed both Doug and Matt to consider what HiFi could and could not do. In Doug's words, "The goal [was] to try not to have too much variation between our tools so that implementation can be relatively consistent among all of our clients." In subsequent meetings, emails, and online discussions, HiFi emerged as a key figure in design choices, with both Matt and Doug questioning whether HiFi could produce particular charts and figures. Even when the Jaxon team moved forward with designs that were outside of HiFi's purview, HiFi's abilities became a hurdle to overcome. For example, the flag icons described above were described as something that HiFi could not create. As a result, the burden fell on Doug to convince Jaxon's Chief Product Officer that its value outweighed the cost of crossing

HiFi's boundaries. Eventually, Doug won this argument, leveraging the client's own appeal to past public feedback.

The continued presence of HiFi in design decisions also shaped Jaxon's interactions with the Carton project team. Because Jaxon was not implementing the final designs, Doug and the Jaxon team could not mention HiFi in their conversations with the Sydney team. As a result, Doug described having to walk a fine line when presenting design rationale to Sydney. When Jaxon's internal design decisions were heavily influenced by appeals to HiFi, Doug would be forced to build alternative narratives to bring to SDOE – in his words, "build[ing] constituency around the design" by appealing to other actors and allies. In the post-project interviews, Doug explained that rather than appealing to HiFi, he instead appealed to Jaxon's experience as a practitioner in the field, convincing SDOE that design choices were based on lessons learned from other clients and other communities of parents.

Interestingly, while technology played a key role in Jaxon's rhetoric, it was never mentioned by the SDOE team. Throughout the project, Sydney's own technology team participated in several design reviews, presumably to vet Jaxon's draft designs and to ensure that Sydney could operationalize the designs with their existing resources. However, no challenges were ever raised. On occasion, members of the Sydney technology team recalled their own struggles in building the state's existing reports; however, these challenges were never extended to the Jaxon designs.

Discussion

The findings above are necessarily limited. Case study research, by design, is limited to looking in-depth to understand the inner workings of a specific environment, without the intention of abstracting that information to make claims about other

environments and other cases. This research, in particular, represents a very specific window into the design process of just two state departments of education working with a single outside contractor. One would expect that the inner workings of accountability report design would vary from state to state, and that different outside contractors might shape their engagements – and the design process – in different ways. Moreover, even within the two cases considered in this article, the findings are necessarily limited by research positionality. I not only conducted the research myself, but also served as a key project team member in one of the two projects, and as an advisor to the second. I have a clear professional relationship with each of the participants involved and I have been deeply involved in shaping Jaxon's approach to design projects. As a result, the data collected and the analytical approach used are necessarily influenced by my position as both practitioner and researcher. That said, within these two cases, there are several noteworthy findings worth additional discussion.

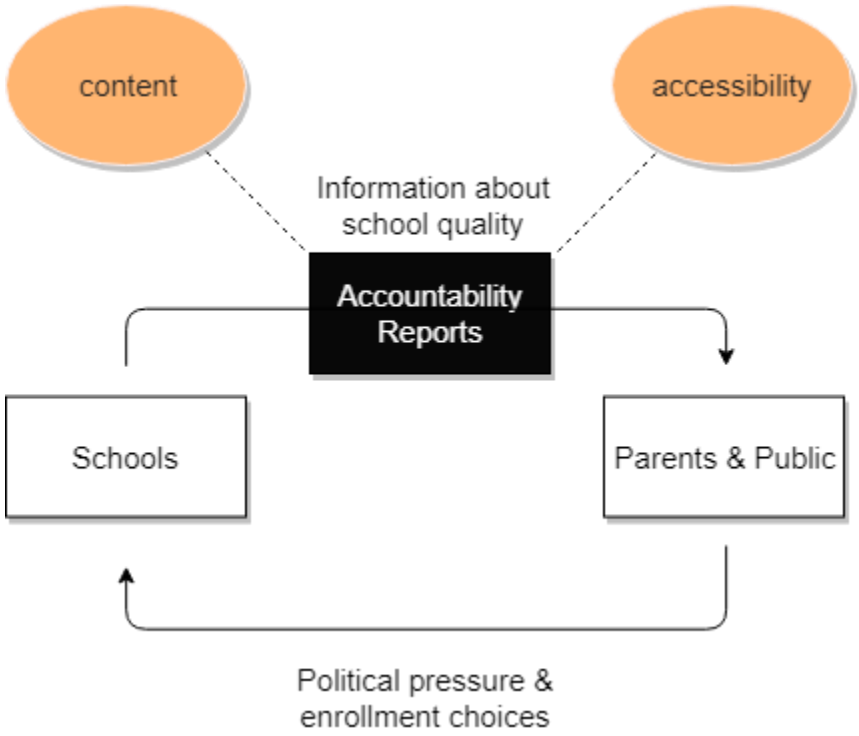
Content Rules

In discussing the design process with members of the Jaxon team, all participants described the primacy of content over form. Before any design decisions were made (and, in fact, before any designers joined the project team), the project managers at Jaxon, CDOE, and SDOE began with a discussion of content, of what information would and would not be reported on these documents. Only after the question of content was largely settled would the designers at Jaxon begin their visualization work.

This is notable because it points to a key tension in the accountability report design process. On the one hand, you have content being mandated at the federal and state level. The federal legislation points to key content areas states must include, and the Department of Education must approve all states' ESSA report content. In addition, state

departments of education are also dictating additional localized content to include on their reports. At the same time, report designers are tasked with the incredibly difficult goal of making clear, accessible reports for non-technical, non-data-savvy parents. However, because content is key, because designers take for granted that their designs must represent any and all mandated content, tensions inevitably arise. When designing accountability reports, one could imagine a world in which content and design are equally considered – in which project teams work to evaluate which content and which design elements, combined, create the most clear, accessible, and coherent picture of school quality. But this is not the case. Instead, form is always second to content, and designers must accommodate content even when it compromises the overall clarity of the reports themselves.

Figure 3. Competing priorities in accountability report design



This tension between content and accessibility threatens the theory of action supporting public school accountability. Accountability reports are intended to provide parents and the public with transparent and unbiased information about school quality so that parents and the public can exert pressure on schools to improve. Yet, as we have seen, the design of these reports has incredible influence over the ways in which audiences interpret and understand information, and therefore, over how they choose to respond to that information. Insofar as the design of accountability reports is always secondary to content, designers will undoubtedly be forced to compromise clarity and accessibility in order to accommodate content.

Presence and Absence of Parents

Another notable finding in the Carton and Sydney cases is the fundamental role played by "parents" despite the absence of any parent's direct participation. As mentioned, many of the key turning points in the design, and many tensions between project stakeholders, revolved around parents. When fighting for *this* design over *that* design, team members defended their positions by asserting their expertise in parents' abilities and desires. The SDOE team pushed for icons, rather than numbers, because that's what parents preferred – after all, they received phone calls from parents stating as much. Similarly, the Jaxon team resolved disputes by embodying parents, asking themselves again and again, what design choices would best meet the needs of parents attempting to quickly understand accountability information.

It is worth noting that although the Carton and Sydney projects did not involve direct parent participation, the Jaxon team did frequently refer to other client projects and independent research that included significant parent involvement, including 1-on-1 interviews, focus groups, and large-scale surveys. Moreover, although Jaxon team

members did not scrutinize their own understanding of parents in the same way they questioned clients, this likely reflects the relevant parties' expertise more than any lack of self-awareness (i.e., departments of education often hire firms like Jaxon precisely because those firms have richer understanding of how to design clear and accessible public reports).

However, the larger takeaway here is that the phenomenon of speaking on behalf of parents is, to some extent, unavoidable. Even when parents are directly and frequently engaged in the design process, it is impossible to solicit feedback from every parent or guardian in the state. Moreover, no matter how many voices are heard, those voices are catalogued, summarized, and repackaged again and again throughout the project. Even in projects with deep qualitative research roots, project stakeholders resolve disputes by asserting their interpretation of parents' feedback, and enrolling parents in defense of one design over another.

This finding further emphasizes how inadequate current legislative guidelines are for supporting accountability report design. As mentioned, the legislation encourages states to solicit parent feedback and to incorporate that feedback throughout the design process, ensuring that design choices reflect parent needs and interests. However, which parent voices will be heard? Who is responsible for speaking on their behalf? In both Carton and Sydney, the needs and desires of parents were certainly taken into account, but it is unclear if the constant enrollment of parents in support (or critique) of design choices actually produced reports that serve parents' needs and interests.

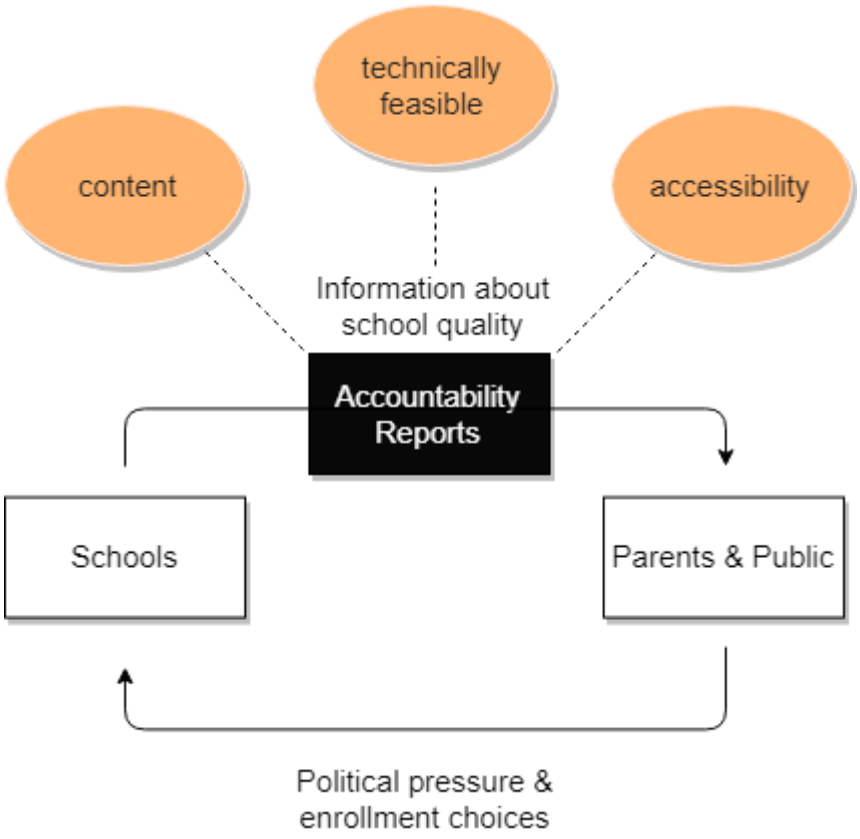
Pulling our perspective back even further, this subjective process of enrolling parents in the design process also complicates the theory of action behind public school accountability. In that model, parents receive clear and transparent information which they can use to push for change. The responsibility of report designers is to create

accountability reports that are intuitive and accessible to the public. However, what does it mean to be intuitive and accessible? In the Sydney and Carton cases, stakeholders continually redefined and reinterpreted these terms by redefining and reinterpreting parents themselves. The abilities and needs of parents existed as a rhetorical construct that could be challenged, renegotiated, and remade. As a result, there is no guarantee that the design accomplished its goal or that the imagined goal of the designers aligned with the pragmatic goals of actual parents and families.

Deference to Tools

In both cases, technology also played a key role in decision-making. In the Carton project, Hillary's assertion that InDesign accommodated one design and not another ultimately resolved the argument over which design to pursue. Similarly, in Sydney, Doug's push to limit designs to align with Jaxon's HiFi software was unquestioned. This enrollment of technology – though used sparingly – was highly successful in resolving disputes and closing the black-box of design. Because of this, technology's role in determining design decisions creates an additional tension in the model of high stakes accountability. Designers must not only build reports that include mandated content and provide a clear, accessible view into that content, they must also design reports that are technically feasible.

Figure 4. Additional design tensions introduced by technology



Complicating things further, the Sydney project emphasizes that this constraint applies even when the technologies are not directly involved in the design process. In Sydney, Jaxon made specific design choices based on the limitations of software, even though the project was never intended to be produced by that software. The implication here, from an ANT theory, is that technology is an incredibly strong actor within the design process, one that stakeholders can enroll in defense of design solutions as easily and effectively as other embodied actors, like parents and the public. Moreover, unlike parents, it is harder for multiple actors to speak on behalf of technology – either one is an expert in a specific technology or one is not. And non-technical actors tend to defer to technical expertise. This deference to technical feasibility surfaces another potential

complication with the model of high stakes accountability: the model assumes that accountability reports are designed to clearly communicate content, but the very ability to communicate content depends on technological systems of production. If and when those systems cannot produce a specific design, the clarity of the design must be compromised to fit within the constraints of the technology. Potentially more troubling, the constraints of the system may become, like the "needs" and "abilities" of parents, a rhetorical construct that various actors negotiate throughout the project, often without consideration to what is and is not possible in practice.

Implications

Jaxon's work with the Carton and Sydney departments of education provides powerful guidelines for future practitioners. Whether these practitioners work for state departments of education or for outside consultants like Jaxon, these cases outline several key concerns for anyone undertaking the difficult work of accountability report design. First, and foremost, the Jaxon cases emphasize the importance of establishing audience. To whom are these reports addressed and with what outcomes in mind? The themes of content before form and of parents as (imagined) actors are both attempts to answer these questions of audience and purpose. To use ANT terminology, the conflicts over what the "black box" should be starts with an assertion of what the black box *does*. Do accountability reports exist merely to check a compliance box? Do they exist to satisfy internal data requests from the districts and schools within the state? Or are they designed primarily to inform parents and families about school quality?

Once establishing these goals, practitioners must be vigilant of attempts to redefine and renegotiate them. These goals and decisions are never as fixed, firm, and straightforward as they appear. Instead, they are continually challenged, reshaped, and

even occasionally rejected throughout the course of the project. With both the Carton and Sydney projects, project team members on all sides constantly reshaped "parents" themselves, whether through role-play (e.g., "If I were a parent…"), reliance on past experience (e.g., "We received several phone calls last year) or even others' assertions (e.g., "Design research shows that parents …"). This constant renegotiation of terms, this redefinition of actors' needs and desires, is a critical moment for future practitioners. As state departments of education embark on updating their ESSA reports, they must pay careful attention to how they and their project teams define parents as an intended audience. This is not to say that project teams should be inflexible, discouraging any competing notions of parents' needs and desires. But, rather, they must be mindful of those moments when anyone speaks on behalf of parents, recognizing that this rhetorical strategy is always an attempt to advocate for one choice over another. By being mindful of these moments and interrogating these moments, project team members can better assess whether a particular claim about parents holds weight.

Similarly, the theme of technology-as-actor provides an additional guideline for state departments of education that are producing accountability reports. Practitioners must realize that report designs are necessarily impacted by the technology used to implement those designs and to produce actual reports. Each production technology has certain affordances and limitations. Practitioners must realize that these technologies may influence design thinking even when those technologies may not be used to produce the reports themselves. In the Sydney case, the Jaxon team described how their design thinking was influenced by the affordances and limitations of their own HiFi software, even though Sydney intended to design the reports themselves. While this may seem self-serving on Jaxon's part, team members explained that these decisions were motivated primarily by concern for their clients. Historically, Jaxon found that many

clients intended to produce accountability reports in-house, but often struggled to do so due to constraints in budget, capacity, and timelines. As a result, these clients would turn to Jaxon, often at the last minute, asking Jaxon to produce the reports instead. To prepare for this eventuality, Jaxon worked to create designs that aligned to their own software whenever possible. That said, this anecdote suggests that practitioners within departments of education might benefit from interrogating the role of technology within the design process. As design decisions are made throughout the duration of the project, practitioners should question whether actors are advocating for one choice over another for technological reasons. Regardless of outcome, asking these questions will help decision-makers better understand trade-offs between designing within technological constraints versus designing to meet stakeholder needs.

Finally, these cases emphasize the need of report designers to move beyond the design process and engage with these reports "in the wild". Future research should consider how the design process changes when parents are invited to speak for themselves, whether through individual participation, interviews, focus groups, or surveys. In these cases, do other actors still assume expertise over parents? Are parents' embodied statements of needs and wants sacrosanct or are they overridden by other actors? Additionally, regardless of whether parents do or do not participate in the design decision, there is a need to better understand how these decisions play out when those designs are put into daily life. The final design of an accountability report is the culmination of numerous choices, many of which are resolved by an appeal to what parents would want to see and how parents would navigate the information displayed. To what extent are these assumptions correct? When parents and families interact with accountability reports, are they doing so in ways that the project team predicted? Are the winning arguments (and the winning designs) having the intended affect? By engaging

with parents directly, researchers can add further depth and nuance to the complicated process of accountability report design, empowering practitioners with better tools in their efforts to achieve ESSA's lofty goals of public school accountability and high quality public education.

CONCLUSION

Ultimately, the findings across these articles call into question the very assumptions of contemporary public school accountability. Public school accountability, as defined by NCLB and ESSA, suggests that when parents and the public are given high quality information about schools, they will hold those schools accountable, and drive positive change. Accountability reports serve as the key mechanism for providing clear and transparent information to parents regarding the quality of education that schools are providing to their students.

Yet, as we have seen, by its very design, accountability legislation complicates its own agenda. Although ESSA calls for states to produce reports that are "concise, presented in an understandable and uniform format, and accessible to the public," the legislation's flexibility with regard to reported content, and its lack of oversight with regard to data designs, has led to accountability reports that rarely achieve this standard. Instead, states have produced accountability reports with incredibly inconsistent designs, a confusing amount of variation from page to page, and an over-reliance on dense tabular displays that overwhelm readers with content. Making matters worse, states often provide no contextual information to help readers make sense of these inconsistent reports – and when they do, the text is often written in complex and indecipherable terms.

Looking into the design process itself, we see why this is the case. As designers work to support administrators in producing accountability reports, they are torn between an unquestioning presentation of mandated content and an appeal towards the actual users of reports: namely, a non-technical audience of parents and the public who are frustrated by technical jargon and dense data displays. Furthermore, in this tug-of-war, we find one side surprisingly absent. While administrators continue to push for

content, parents are rarely directly engaged in the design process and instead serve as a hypothetical actor whose needs and desires are taken up by various other stakeholders throughout the design process.

Moving forward, these findings should serve as a clarion call for policymakers, practitioners, and researchers alike. Given the importance of accountability reports as a lever of school accountability, there is a disappointing lack of existing research on the design and dissemination of these reports. Future research is needed to better understand the relationship between policymakers' intentions, the laws that they enact, and the actions that state education administrators take in response. Moreover, additional research is needed to bridge the gap between accountability report design and audiences' engagement with the reports themselves. More work must be done to understand how parents do (and do not) make sense of current accountability reports so that we can make stronger strides towards the ideal of giving families clear and transparent information about the quality of education provided their public schools.

APPENDIX A: INTERVIEW PROTOCOL

Interview Protocol
Production of School Accountability Reports at the State Level
Design Consultant Interview Draft

Interviewee:	Last, First
Interviewer:	Moore, Michael
Location:	(TBD)

Participant Interview Preparation
None

Interviewer Directions
This document serves as a rough guide for conducting the interview. To preserve candid and earnest conversation, questions may be asked out of order. Interviewers are encouraged to probe for elaboration.

Project Background
The purpose of this study is to better understand how specific design decisions were made in creating the [STATE] Department of Education's school accountability reports. In particular, we are interested in asking participants to put concrete design artifacts, as well as changes to those artifacts over time, into the larger context of interpersonal interactions during their experience with the project.

Interview Introduction
Thank you so much for taking the time to talk today. As I mentioned when reaching out to you, my dissertation research is on the development of state-level school accountability reports. In particular, I am interested in getting your opinion on how specific design choices, (for example how to chart student achievement or graduation rates,) were made.

There are no risks to participating in this project. All of your responses in the interview will be kept completely confidential. The project name, as well as the name of the individuals, organizations, and states involved will be changed to preserve anonymity. This interview will take no more than 60 minutes.

Do you have any questions for me?

Is it OK with you if I record this interview? The recording will be for myself only, to ensure I have accurately captured your responses. All recorded material will be destroyed after use.

Interview Questions

Background to Project

I want to begin by asking a few broad background questions about your role in the School Accountability Report Card (SSARC) project between [COMPANY] and [STATE].

1. What was your job title through the duration of this project?

2. What were the primary functions of your role at [COMPANY]?

3. What were your specific responsibilities during the SSARC project?

4. In reflecting on the SSARC project, what do you believe had the greatest impact on the final design of the published PDF reports?
    a. Probe: What evidence leads you to this conclusion? (Specific examples)

Using Artifacts to Understand Actor Networks

In these next few questions, I want to call your attention to the reports themselves. However, before we start, I want to ask:

5. Over the course of the project, can you remember any moments where the design changed from one draft to the next?
    a. Probe: Push for concrete example.
    b. Probe: Within concrete example: what do you believe led to this change?
    c. Probe: Beyond concrete example: did this happen frequently?

With that in mind, here are several design drafts that I have collected. Each draft represents the design of the SSARC reports at various stages throughout the course of the project. Moreover, these draft designs are paired. In each pair

6. For any of the drafts in front of you, do you remember this change occurring? If so, what do you believe led to this change?
    a. Probe: Encourage discussion of multiple pairs of drafts
    b. Probe: If answers vary from question 4 above, interrogate.

Although our discussion is based on the [STATE] SSARC reports, many other states report similar data on their own accountability reports. In my dissertation research I have found [metric] is reported by many states, but often in different ways.

7. Look at [metric] on the SSARC drafts in front of you. Can you recall how this design came to be?
    a. Probe: Was the choice consistent from the start, or did it vary?

Interpretive Flexibility

Before we finish, I'd like to switch gears. Instead of reflecting on what *did* happen with the SSARC, I'd like to ask a hypothetical question:

8. Based on your experience, can you imagine a world in which the final report design contained the exact same content, but presented in a different format? If so, how would this most likely have happened? If not, why not?

Probe: What areas / visualizations seem most “flexible”?

REFERENCES

Aschbacher, P. R., & Herman, J. L. (1991). *Guidelines for Effective Score Reporting* (CSE Technical Report No. 326). Center for Research on Evaluation, Standards, and Student Testing: UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/TR326.pdf

Balchin, W. G. V. (1972). Graphicacy. *Geography*, *57*(3), 185–195.

Benbasat, I., & Dexter, A. S. (1986). An Investigation of the Effectiveness of Color and Graphical Information Presentation under Varying Time Constraints. *MIS Quarterly*, *10*(1), 59–83. https://doi.org/10.2307/248881

Bendick, M., & Cantú, M. G. (1978). The Literacy of Welfare Clients. *Social Service Review*, *52*(1), 56–68.

Beniger, J. R., & Robyn, D. L. (1978). Quantitative Graphics in Statistics: A Brief History. *The American Statistician*, *32*(1), 1–11. https://doi.org/10.1080/00031305.1978.10479235

Bijker, W., Hughes, T. P., & Pinch, T. (1989). *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. The MIT Press.

Boston, C. (2002). The Concept of Formative Assessment. ERIC Digest. Retrieved from http://eric.ed.gov/?id=ED470206

Breland, K., & Breland, M. (1944). Legibility of newspaper headlines printed in capitals and in lower case. *Journal of Applied Psychology*, *28*(2), 117–120. https://doi.org/10.1037/h0053815

Brinton, W. C. (1914). *Graphic methods for presenting facts*. New York: The Engineering Magazine Company.

Bush, G. W. (2001, August). *President Discusses Education at National Urban League Conference*.

Callon, M. (1984). Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. *The Sociological Review*, *32*, 196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Carnoy, M., Elmore, R., & Siskin, L. (Eds.). (2003). *The New Accountability: High Schools and High-Stakes Testing* (1 edition). New York: Routledge.

Cleveland, W. S. (1994). *The Elements of Graphing Data* (2 edition). Murray Hill, N.J: Hobart Press.

Cleveland, W. S., & McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the*

American Statistical Association, *79*(387), 531–554. https://doi.org/10.1080/01621459.1984.10478080

Coleman, M., & Liau, T. L. (1975). A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, *60*(2), 283–284.

Cooley, M. E., Moriarty, H., Berger, M. S., Selm-Orr, D., Coyle, B., & Short, T. (1995). Patient literacy and the readability of written cancer educational materials. *Oncology Nursing Forum*, *22*(9), 1345–1351.

Creswell, J. W. (1997). *Qualitative Inquiry and Research Design: Choosing among Five Traditions* (1st ed.). Sage Publications, Inc.

Darling-Hammond, L. (2006). No Child Left Behind and High School Reform. *Harvard Educational Review*, *76*(4), 642–667. https://doi.org/10.17763/haer.76.4.d8277u8778245404

Doak, C., Doak, L., & Root, J. (1995). *Teaching Patients with Low Literacy Skills* (Second edition). Philadelphia: LWW.

Elementary and Secondary Education Act of 1965: H.R. 2362, Pub. L. No. Public Law 89-10 (1965).

Ellul, J. (1964). The technological society ([1st American ed.]). New York: Knopf.

Eltorai, A. E. M., Sharma, P., Wang, J., & Daniels, A. H. (2015). Most American Academy of Orthopaedic Surgeons' Online Patient Education Material Exceeds Average Patient Reading Level. *Clinical Orthopaedics and Related Research®*, *473*(4), 1181–1186. https://doi.org/10.1007/s11999-014-4071-2

Every Student Succeeds Act of 2015, Pub. L. No. 114–95, § S. 1117 (2015). Retrieved from https://www.congress.gov/bill/114th-congress/senate-bill/1177/text

Every Student Succeeds Act State and Local Report Cards Non-Regulatory Guidance. (2017). Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/essastatereportcard.pdf

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Forte, E. (2010). Examining the Assumptions Underlying the NCLB Federal Accountability Policy on School Improvement. *Educational Psychologist*, *45*(2), 76–88. https://doi.org/10.1080/00461521003704738

Foster, J., & Coles, P. (1977). An Experimental Study of Typographic Cueing in Printed Text. *Ergonomics*, *20*(1), 57–66. https://doi.org/10.1080/00140137708931601

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145–220.

Gribbons, W. M. (1992). Organization by Design: Some Implications for Structuring Information. *Journal of Technical Writing and Communication*, *22*(1), 1–1. https://doi.org/10.2190/9BD5-3QFX-PB7J-EBH2

Hambleton, R. K., & Slater, S. C. (1996). Are NAEP Executive Summary Reports Understandable to Policy Makers and Educators?. Retrieved from http://eric.ed.gov/?id=ED400296

Hansberry, D. R., Agarwal, N., Gonzales, S. F., & Baker, S. R. (2014). Are We Effectively Informing Patients? A Quantitative Analysis of On-Line Patient Education Resources from the American Society of Neuroradiology. *American Journal of Neuroradiology*, *35*(7), 1270–1275. https://doi.org/10.3174/ajnr.A3854

Hanushek, E. A., & Raymond, M. E. (2005). Does School Accountability Lead to Improved Student Performance? *Journal of Policy Analysis and Management*, *24*(2), 297–327.

Hastings, J. S., & Weinstein, J. M. (2008). Information, School Choice, and Academic Achievement: Evidence from Two Experiments. *Quarterly Journal of Economics*, *123*(4), 1373–1414.

Hattie, J. (2009). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal*.

Hess, F. M., & Petrilli, M. J. (2007). *No Child Left Behind Primer* (2 edition). New York: Peter Lang International Academic Publishers.

Horn, R. (2002). *Understanding Educational Reform: A Reference Handbook* (annotated edition edition). Santa Barbara, Calif: ABC-CLIO.

Horton, W. (1991). Overcoming chromophobia: A guide to the confident and appropriate use of color. *Professional Communication, IEEE Transactions On*, *34*(3), 160–171.

Huff, D. (1993). *How to Lie with Statistics* (Reissue edition). New York: W. W. Norton & Company.

Isaacs, M. L. (2003). Data-Driven Decision Making: The Engine of Accountability. *Professional School Counseling*, *6*(4), 288–295.

Jacobsen, R., Snyder, J. W., & Saultz, A. (2014). Informing or Shaping Public Opinion? The Influence of School Accountability Data Format on Public Perceptions of School Quality. *American Journal of Education*, *121*(1), 1–27.

Jarvenpaa, S. L., & Dickson, G. W. (1988). Graphics and Managerial Decision Making: Research-based Guidelines. *Commun. ACM*, *31*(6), 764–774. https://doi.org/10.1145/62959.62971

Kane, T. J., & Staiger, D. O. (2001). *Improving School Accountability Measures* (Working Paper No. 8156). National Bureau of Economic Research. https://doi.org/10.3386/w8156

Kane, T. J., & Staiger, D. O. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *The Journal of Economic Perspectives*, *16*(4), 91–114.

Karsten, K. G. (1923). *Charts and Graphs*. Prentice Hall.

Katzir, T., Hershko, S., & Halamish, V. (2013). The Effect of Font Size on Reading Comprehension on Second and Fifth Grade Children: Bigger Is Not Always Better. *PLOS ONE*, *8*(9), e74061. https://doi.org/10.1371/journal.pone.0074061

Kelly, P. A., & Haidet, P. (2007). Physician overestimation of patient literacy: A potential source of health care disparities. *Patient Education and Counseling*, *66*(1), 119–122. https://doi.org/10.1016/j.pec.2006.10.007

Kincaid, J. P., Fishburne, J., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel* (No. RBR-8-75). NAVAL TECHNICAL TRAINING COMMAND MILLINGTON TN RESEARCH BRANCH, NAVAL TECHNICAL TRAINING COMMAND MILLINGTON TN RESEARCH BRANCH. Retrieved from http://www.dtic.mil/docs/citations/ADA006655

Kirby, S. N., McCaffrey, D. F., Lockwood, J. R., McCombs, J. S., Naftel, S., & Barney, H. (2002). Using State School Accountability Data to Evaluate Federal Programs: A Long Uphill Road. *Peabody Journal of Education*, *77*(4), 122–145.

Klein, A. (2016). Under ESSA, states, districts to share more power. *Education Week*, *35*(15), 10–12.

Krippendorff, K. H. (2012). *Content Analysis: An Introduction to Its Methodology* (Third Edition edition). Los Angeles ; London: SAGE Publications, Inc.

Ladd, H. F. (2001). School—Based Educational Accountability Systems: The Promise and the Pitfalls. *National Tax Journal*, *54*(2), 385–400.

Latour, B. (1988). *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369–381.

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory* (1st edition). Oxford University Press.

Learning Heroes. (2017). *Parents 2017: Unleashing Their Power & Potential*. Learning Heroes.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability Systems: Implications of Requirements of the No Child Left behind Act of 2001. *Educational Researcher*, *31*(6), 3–16.

Macdonald-Ross, M. (1977). How numbers are shown. *AV Communication Review*, *25*(4), 359–409. https://doi.org/10.1007/BF02769746

McGuinn, P. (2016). From No Child Left behind to the Every Student Succeeds Act: Federalism and the Education Legacy of the Obama Administration. *Publius: The Journal of Federalism*, *46*(3), 392–415. https://doi.org/10.1093/publius/pjw014

McMurrer, J., & Yoshioka, N. (2013). *States' Perspectives on Waivers: Relief from NCLB, Concern about Long-Term Solutions*. Center on Education Policy. Retrieved from https://eric.ed.gov/?q=two+state+solution+for+Israel&ff1=subFederal+Legislation&id=ED555343

McNeil, M., & Klein, A. (2011). Obama Outlines NCLB Flexibility. *Education Week*, *31*(5).

Meade, C. D., & Byrd, J. C. (1989). Patient literacy and the readability of smoking education literature. *American Journal of Public Health*, *79*(2), 204–206. https://doi.org/10.2105/AJPH.79.2.204

Mikulecky, M., & Christie, K. (2014). *Rating States, Grading Schools*. Education Commission of the States. Retrieved from http://www.ecs.org/docs/rating-states,grading-schools.pdf

Mumford, L. (1934). Technics and civilization. New York: Harcourt, Brace and company.

Nagro, S. A., & Stein, M. L. (2016). Measuring Accessibility of Written Communication for Parents of Students With Disabilities: Reviewing 30 Years of Readability Research. *Journal of Disability Policy Studies*, *27*(1), 13–21. https://doi.org/10.1177/1044207314557489

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 20 U.S.C. (2001).

Poracsky, J., Young, E., & Patton, J. P. (1999). The Emergence of Graphicacy. *The Journal of General Education*, *48*(2), 103–110.

Powers, R. D. (1988). Emergency department patient literacy and the readability of patient-directed materials. *Annals of Emergency Medicine*, *17*(2), 124–126. https://doi.org/10.1016/S0196-0644(88)80295-6

Ravitch, D. (2010). *The Death and Life of the Great American School System: How Testing and Choice Are Undermining Education*. New York: Basic Books.

Roberts, M. R., & Gierl, M. J. (2009, April). *Development of a Framework for Diagnostic Score Reporting*. Presented at the Annual meeting of the American Educational Research Association, San Diego, CA. Retrieved from http://www2.education.ualberta.ca/educ/psych/crame/files/AERA%202009%20diagnostic_score_reporting%20Mar%203.pdf

Rogers, J. (2006). Forces of Accountability? The Power of Poor Parents in NCLB. *Harvard Educational Review*, *76*(4), 611–641. https://doi.org/10.17763/haer.76.4.846v832864v51028

Rose, L. C. (2004). No Child Left Behind: The Mathematics of Guaranteed Failure. *Educational Horizons*, *82*(2), 121–130.

Schonlau, M., & Peters, E. (2012). Comprehension of Graphs and Tables Depend on the Task: Empirical Evidence from Two Web-Based Studies. *Statistics, Politics, and Policy*, *3*(2). https://doi.org/10.1515/2151-7509.1054

Schriver, K. A. (1997). *Dynamics in document design: creating text for readers*. John Wiley & Sons, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=249331

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, *91*(4), 690–702. https://doi.org/10.1037/0022-0663.91.4.690

Shaul, M. S., & Ganson, H. C. (2005). The No Child Left behind Act of 2001: The Federal Government's Role in Strengthening Accountability for Student Performance. *Review of Research in Education*, *29*, 151–165.

Sirotnik, K. A. (2004). *Holding Accountability Accountable: What Ought to Matter in Public Education*. New York ; London: Teachers College Press.

Smith, M. R., & Marx, L. (1994). *Does Technology Drive History?: The Dilemma of Technological Determinism*. MIT Press.

Snyder, T. D., & Dillow, S. A. (2014). *Digest of Education Statistics 2012*. Government Printing Office.

Spillane, J. P. (2012). Data in Practice: Conceptualizing the Data-Based Decision-Making Phenomena. *American Journal of Education*, *118*(2), 113–141. https://doi.org/10.1086/663283

State and Local Report Cards: Title I, Part A of the Elementary and Secondary Education Act of 1965, as Amended, Non-Regulatory Guidance. (2013, February 8). U.S. Department of Education. Retrieved from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Stewart, B. M., Cipolla, J. M., & Best, L. A. (2009). Extraneous information and graph comprehension: implications for effective design choices. *Campus-Wide Information Systems*, *26*(3), 191–200.

Thomsen, J. (2013). *50-State Comparison: State School Accountability Report Cards*. Education Commission of the States. Retrieved from http://www.ecs.org/state-school-accountability-report-cards/

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, Conn.: Graphics Pr.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd edition). Cheshire, Conn: Graphics Pr.

Tufte, E. R. (2006). *Beautiful Evidence* (1St Edition edition). Cheshire, Conn: Graphics Pr.

Tukey, J. W. (1990). Data-Based Graphics: Visual Display in the Decades to Come. *Statistical Science*, *5*(3), 327–339. https://doi.org/10.1214/ss/1177012101

Vaiana, M. E., & McGlynn, E. A. (2002). What Cognitive Science Tells Us about the Design of Reports for Consumers. *Medical Care Research and Review*, *59*(1), 3–35. https://doi.org/10.1177/107755870205900101

Wainer, H. (1997). Improving Tabular Displays, With NAEP Tables as Examples and Inspirations. *Journal of Educational and Behavioral Statistics*, *22*(1), 1–30. https://doi.org/10.3102/10769986022001001

Wiliam, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist*, *45*(2), 107–122. https://doi.org/10.1080/00461521003703060

Winn, W. (1991). Color in document design. *Professional Communication, IEEE Transactions On*, *34*(3), 180–185.

Winner, L. (1988). *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University Of Chicago Press.

Wong, K. K. (2008). Federalism Revised: The Promise and Challenge of the No Child Left Behind Act. *Public Administration Review*, *68*, S175–S185.

Wong, K. K. (2015). Federal ESEA Waivers as Reform Leverage: Politics and Variation in State Implementation. *Publius: The Journal of Federalism*, *45*(3), 405–426. https://doi.org/10.1093/publius/pjv020

Yau, N. (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. John Wiley & Sons.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116–138.